4-13-2018 2:00 PM

Exact Box-Cox Analysis

Samira Soleymani
The University of Western Ontario

Supervisor
Dr. A. Ian McLeod
The University of Western Ontario

Graduate Program in Statistics and Actuarial Sciences
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

The Box-Cox method has been widely used to improve estimation accuracy in different fields, especially in econometrics and time series. In this thesis, we initially review the Box-Cox transformation [Box and Cox, 1964] and other alternative parametric power transformations. Following, the maximum likelihood method for the Box-Cox transformation is presented by discussing the problems of previous approaches in the literature.

This work consists of the exact analysis of Box-Cox transformation taking into account the truncation effect in the transformed domain. We introduce a new family of distributions for the Box-Cox transformation in the original and transformed data scales. A likelihood analysis of the Box-Cox distribution is presented when truncation is considered. It is shown that numerical problems may arise in prediction and simulation when the truncation effect is ignored.

A new algorithm has been developed for simulating Box-Cox transformed time series since previous methods are inefficient or unreliable. An application to sunspot data is discussed.

Box-Cox analysis is employed for random forest regression prediction using cross-validation instead of MLE to estimate the transformation. An application to Boston housing dataset demonstrates that this technique can substantially improve prediction accuracy.

Keywords: Box-Cox transformation, cross-validation, maximum likelihood, time series simulation, truncated distributions

Dedicated to my parents for their love, support and encouragement.

Acknowledgements

This thesis could not have been accomplished without the help of many people, to whom I am truly indebted for their valuable contributions. Through theses years of PhD, living in this warm family at Department of Statistical and Actuarial Sciences at Western University, I gained a big family where I can get energy to conquer all the future obstacles.

First and foremost I would like to express my sincere gratitude to my supervisor Dr. A. Ian McLeod for his invaluable guidance and generous support throughout my graduate experience at Western University. Without his encourage I would never accomplish the completion of my Ph.D journey. I am also grateful to all faculty, staff and fellow students at the Department of Statistical and Actuarial Sciences for their encouragement. Special thanks are also devoted to my examiners, Dr. Serge Provost, Dr. Sudhir Paul, Dr. Neil Klar and Dr. Jiandong Ren for their insightful comments and suggestions.

I am also indebted to my instructors at Western University, including but not limited to Dr. Wenqing He, Dr. Reg Kulperger, Dr. Hao Yu. I have the opportunity to work with, Dr. Duncan Murdoch, Dr. Hristo Sendov and Dr. Xiaoming Liu as their teaching assistant, and also Dr. David Bellhouse and Dr. Bethany White as their statistical consultant. Their encouragement and assistance are highly appreciated and will always be remembered.

Last but foremost, I would like to thank my father, Behzad, and mother, Nasrin for their perpetual and unconditional love and support through my life. I also would like to say thanks to my husband, Masoud and sister, Simin.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter a review is presented regarding the parametric power transformations in regression models and time series. Box and Cox [1964] proposed the Box-Cox transformation in order to improve the statistical models. It has been extensively studied on this subject with the most of the research concentrated on inferences about unknown parameters of interest [Box and Cox, 1964, Bickel and Doksum, 1981, Hinkley and Runger, 1984, Carroll and Ruppert, 1981].

In the literature, it was assumed that parametric family of distribution from y to $y^{(\lambda)}$ with λ parameter are normally distributed with constant variance σ^2 and mean μ. Therefore, the probability density for inverse-transformed observations and likelihood function in original domain were obtained by multiplying the normal density distribution by the Jacobian of transformation. The violation of the assumptions was sometimes ignored and statistical analysis was performed even though all assumptions were not satisfied.

In Chapter 2, we investigate the Box-Cox family distribution in the untransformed data domain by considering the truncation effect. Our next objective is to find that the optimal transformation would be changed by this assumption and how log-likelihood would be differed with respect to parameters λ, μ and σ. The exact Box-Cox distribution is shown and the comparison between an approximate and exact analysis is discussed in details. One of the controversial question would be regarding the estimation of λ, and also it needs to explore that MLE would work under this assumption and limitation.

Chapter 2 explors the contribution of the exact Box-Cox analysis with application to rivers dataset. Initially, Chen and Lockhart [1997] obtained the conditional and unconditional inferences using parameter-based asymptotics for large finite sample size. Further, Chen et al. [2002] employed a large sample theory in the Box-Cox linear models and developed the goodness-of-fit test for the Box-Cox transformation. We illustrate the findings of Chapter 2 with simulated examples in a regression model, and compare them with some of the work done by previous researchers.

Finally, we revisit carefully the simulated truncated normal by Robert [1995], and we propose an efficient algorithm for generating the Box-Cox transformed time series. It is presented an example of the truncation problem with Box-Cox analysis. Further, the modifications are performed for Durbin-Levinson algorithm to improve the simulation of the Box-Cox time series.

This thesis reviews the transformation effect in linear regression, time series and Machine

Learning. The simple power transformation is employed to minimize the expected prediction error.

1.1 Review of the Box-Cox Transformation

Transformation used to stabilize variance if variance changes with mean level of measurements [Bartlett, 1947]. Tukey [1957] introduced the power transformation to achieve normality of distribution or at least symmetrizing error distribution. This transformation is monotone and it preserves the order of data for $\lambda > 0$. Power transformations can be defined for positive random variable as,

$$Y^{(\lambda)} = \begin{cases} Y^{\lambda}, & \text{if } \lambda \neq 0, \\ \log(Y), & \text{if } \lambda = 0. \end{cases} \tag{1.1}$$

Power transformations are often used in econometric and Kriging applications. The usual practice is to back transform data into the original data domain. Box and Tidwell [1962] concentrated on the transformation of independent variables with no impact on homoscedasticity and normalization of error distribution. The Box-Cox transformation is a linear transformation of the power transformation but it is more suitable for mathematical treatment. For any $Y > 0$ and $\lambda \epsilon R$, the Box-Cox transformation is given by,

$$Y^{(\lambda)} = \begin{cases} (Y^{\lambda} - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log(Y), & \text{if } \lambda = 0. \end{cases} \tag{1.2}$$

Box and Cox [1964] suggested the Box-Cox transformation which can improve statistical models by,

1. removing non-linearity;
2. removing heteroscedasticity;
3. removing skewness and non-normality of errors.

The inverse Box-Cox transformation is obtained by,

$$Y = \begin{cases} (\lambda Y^{(\lambda)} + 1)^{1/\lambda}, & \text{if } \lambda \neq 0, \\ \exp(Y^{(\lambda)}), & \text{if } \lambda = 0. \end{cases} \tag{1.3}$$

Shifted power transformation was proposed to handle the negative observation [Box and Cox, 1964] as,

$$Y^{(\lambda)} = \begin{cases} ((Y + \lambda_2)^{\lambda_1} - 1)/\lambda_1, & \text{if } \lambda_1 \neq 0, \\ \log(Y + \lambda_2), & \text{if } \lambda_1 = 0. \end{cases} \tag{1.4}$$

where λ_1 is the transformation parameter and λ_2 is shifted parameter defined for $Y + \lambda_2 > 0$. This problem can be considered as non-regular case due to the restriction $Y > -\lambda_2$. The log-likelihood may not be determined by using the two parameter transformations since it approximately tends to infinity as $\lambda_2 + min(Y) \to 0$. Bickel and Doksum [1981] discussed the signed power transformation such that,

$$Y^{(\lambda)} = \{\text{sgn(Y)}|Y|^{\lambda} - 1)\}/\lambda, \quad \text{for} \quad \lambda > 0. \tag{1.5}$$

which can cover the whole real number and there is no restriction on $Y^{(\lambda)}$. This transformation only can address the kurtosis rather than skewness of distribution. The disadvantage of the signed power transformation is that it can not handle the skewed distribution. The initial assumption defined by Box and Cox [1964] restricted to positive data. Yeo and Johnson [2000] generalized the Box-Cox transformation in the case of the negative random variable.
The Yeo-Johnson transformation for a fixed λ, $Y^{(\lambda)}$: $R \to R$ is defined by,

$$Y^{(\lambda)} = Y^{(\lambda)}(\lambda, Y) = \begin{cases} \left((Y+1)^{\lambda} - 1\right)/\lambda, & \text{if } Y \geq 0, \lambda \neq 0, \\ \log(Y+1), & \text{if } Y \geq 0, \lambda = 0, \\ -\left((1-Y)^{2-\lambda} - 1\right)/(2-\lambda), & \text{if } Y < 0, \lambda \neq 2, \\ -\log(1-Y), & \text{if } Y < 0, \lambda = 2. \end{cases} \tag{1.6}$$

where λ is power parameter likewise the Box-Cox transformation. This transformation can hold the properties of the log-mean standardization after the inverse-transformation since $Y^{(\lambda)}$ is invertible. Back-transformation is obtained by,

$$Y = \begin{cases} \left(Y^{(\lambda)}(\lambda, Y)\lambda + 1\right)^{1/\lambda} - 1, & \text{if } Y^{(\lambda)}(\lambda, Y), \geq 0, \lambda \neq 0, \\ \exp(Y^{(\lambda)}(\lambda, Y)) - 1, & \text{if } Y^{(\lambda)}(\lambda, Y) \geq 0, \lambda = 0, \\ 1 - \left(-Y^{(\lambda)}(\lambda, Y)(2-\lambda) + 1\right)^{1/(2-\lambda)}, & \text{if } Y^{(\lambda)}(\lambda, Y) < 0, \lambda \neq 2, \\ 1 - \exp\left(-Y^{(\lambda)}(\lambda, Y)\right), & \text{if } Y^{(\lambda)}(\lambda, Y) < 0, \lambda = 2. \end{cases} \tag{1.7}$$

Lemma 1.1.1 *From the Yeo-Johnson transformation, we can conclude the following results:*

1. For $Y \geq 0$, we have $Y^{(\lambda)}(\lambda, Y) \geq 0$, and for $Y < 0$, it becomes $Y^{(\lambda)}(\lambda, Y) < 0$;
2. $Y^{(\lambda)}(\lambda, Y)$ is continuous function in terms of λ and Y;
3. $Y^{(\lambda)}(\lambda, Y)$ is convex with $\lambda > 1$, and concave with $\lambda < 1$.

1.2 Estimation of the Transformation Parameter

Box and Cox [1964] presented maximum likelihood and Bayesian approach for the estimation of the parameter λ. Maximum-likelihood estimates are obtained for a fixed λ by ignoring constant part as follows [Box and Cox, 1964],

$$\log L_{max}(\lambda) = -\frac{n}{2}\log(\hat{\sigma}^2(\lambda)) + \log J(\lambda; y). \tag{1.8}$$

The robustness of the estimates of the parameters in a linear regression model have been extensively studied and the truncation effect was neglected by Draper and Cox [1969], Atkinson [1973], Bickel and Doksum [1981], Carroll [1980], Hinkley and Runger [1984], Carroll and Ruppert [1981] and Taylor [1986]. Bickel and Doksum [1981] discussed the consistency of parameters via maximum likelihood estimation (MLE) and the asymptotic variance of these estimate in the regression. The ordinary likelihood function may be poorly behaved when the range of observations depends on unknown parameter or no local maximum found [Atkinson and Pericchi, 1991].

Bickel and Doksum [1981] assumed that $Y^{(\lambda)} = h(Y, \lambda)$ can be defined as a monotone increasing transformation of Y in terms of λ in specific interval. Then, h have partial derivation in terms of Y

$$J(Y, \lambda) = \prod_{i=1}^{n} h_Y(Y_i, \lambda), \tag{1.9}$$

where the Jacobian shown the mapping the $(Y_1, ..., Y_n) \rightarrow (h(Y_1, \lambda), ..., h(Y_n, \lambda))$ and the mean of the transformed variable can be written as,

$$\mu_i(\beta) = \sum_{j=1}^{p} x_{ij}\beta_j, \quad \mu_i^0 = \mu_i(\beta_0). \tag{1.10}$$

The likelihood function is

$$L(Y, \beta, \sigma, \lambda) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f\left(\frac{Y_i^{(\lambda)} - \mu_i(\beta)}{\sigma}\right) J(\lambda, Y). \tag{1.11}$$

The main differences between the model by Bickel and Doksum [1981] and Box and Cox [1964] is the use of arbitrary f instead of the normal distribution.

Inference discussion regarding the regression parameter β and λ were presented in different cases [Bickel and Doksum, 1981, Hinkley and Runger, 1984, Carroll and Ruppert, 1981]. Asymptotic calculations were presented such that the estimation of β is asymptotically more variable compare to standard linear model approach, therefore the changes of variance could be significant [Bickel and Doksum, 1981]. Carroll and Ruppert [1981] indicated that the inverse transformation used in order to make efficient inferences in original scale domain. Bickel and Doksum [1981] advocated that maximum likelihood estimates are sensitive to the distribution assumption. As a result, they concluded that it would be crucial to make the normal assumption of the response variable in linear model for inferences about regression coefficients.

The Box-Cox transformation assumes that the variable to be transformed is positive. Thus, in terms of definition, the Box-Cox transformation is intended to induce a truncated normal distribution. Poirier [1978] stated that it would be hard to determine whether the truncation effect is negligible since it depends on the unknown parameters of the distribution including the Box-Cox parameter λ. The Box-Cox transformation used in limited dependent variable (LDV) models with skewness for variables which have likely been censored or truncated [Poirier, 1978]. The estimation approach is to maximize the likelihood function of the truncated normal distribution.

Draper and Cox [1969] have shown that if a power transformation satisfy non-normality, the transformation estimated would be approximately robust corresponding to a distribution

nearly symmetrical distribution and it can be useful. Carroll [1980] suggested a new method to obtain robust estimator rather than likelihood method. Furthermore, approximate normality was investigated in theory and Monte-Carlo approach implemented in linear model.

Chen and Lockhart [1997] argued that the variances of parameter estimators increase based on $\hat{\lambda}$ and parameters are correlated. They derived the Fisher information matrix and its inverse for a general model involving regression as the truncation part was ignored. It was suggested that it can be considered the effect of truncation, but the analysis of likelihood function would be controversial. The models mentioned for the transformed and untransformed Box-Cox distributions can be generalized by using exponential-family distribution as the error distribution.

Suppose $Y_i^{(\lambda)}$ are positive and independent variables with probability density function $\phi(.)$ and cumulative distribution function $\Phi(.)$ of standard normal. Moreover, we define $\xi = -(\lambda^{-1} + \mu)/\sigma$ as a truncation point and $Y_i^{(\lambda)}$ are the transformed variables. Let $\dot{Y}_i^{(\lambda)}$ and $\ddot{Y}_i^{(\lambda)}$ be the first and second derivatives of $Y_i^{(\lambda)}$ with respect to λ respectively. The log-likelihood function in terms of the untransformed variable when truncation is negligible can be written,

$$\log L(\sigma, \mu, \lambda) = -n/2 \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i^{(\lambda)} - \mu_i\right)^2 + (\lambda - 1) \sum_{i=1}^{n} \log y_i. \tag{1.12}$$

In general, the Fisher information matrix for observation Y_i by parameters $\theta = (\sigma, \mu, \beta, \lambda)$ obtained by,

$$1/n \sum_{i=1}^{n} \sigma^2 I_i = \begin{bmatrix} 2 & 0 & 0 & -2a \\ 0 & 1 & 0 & -b \\ 0 & 0 & Q & -C \\ -2a & -b & -C^T & d \end{bmatrix} \tag{1.13}$$

where

$$a = 1/n \sum_{i=1}^{n} E\left[\frac{(Y_i^{(\lambda)} - \mu_i)\dot{Y}_i^{(\lambda)}}{\sigma}\right], \tag{1.14}$$

$$b = 1/n \sum_{i=1}^{n} E\left(\dot{Y}_i^{(\lambda)}\right), \tag{1.15}$$

$$C = 1/n \sum_{i=1}^{n} E[\dot{Y}_i^{(\lambda)} x_i], \tag{1.16}$$

$$d = 1/n \sum_{i=1}^{n} E\left[(Y_i^{(\lambda)} - \mu_i)\ddot{Y}_i^{(\lambda)} + (\dot{Y}_i^{(\lambda)})^2\right], \tag{1.17}$$

$$Q = 1/n X^T X. \tag{1.18}$$

The average information matrix is presented by a, b, C, Q which are dependent on a sample size of n. We need to compute the inverse of Fisher information matrix to determine the asymptotic

variance of MLE of θ. The inverse of the average information matrix is presented by,

$$\Sigma = \left(1/n \sum_{i=1}^{n} I_i\right)^{-1} = \begin{bmatrix} \frac{1}{2} + \frac{a^2}{f} & \frac{ab}{f} & \frac{aC^TQ^{-1}}{f} & \frac{a}{f} \\ \frac{ab}{f} & 1 + \frac{b^2}{f} & \frac{bC^TQ^{-1}}{f} & \frac{b}{f} \\ \frac{aQ^{-1}C}{f} & \frac{bQ^{-1}C}{f} & Q^{-1} + \frac{Q^{-1}CC^TQ^{-1}}{f} & \frac{Q^{-1}C}{f} \\ \frac{a}{f} & \frac{b}{f} & \frac{C^TQ^{-1}}{f} & \frac{1}{f} \end{bmatrix} \tag{1.19}$$

where $d - 2a^2 - b^2 - C^TQ^{-1}C = f$. It will be shown how to make inference about parameters when λ is unknown. Under regularity conditions, we can show that the distribution of $\hat{\theta}$ can be approximated by,

$$\sqrt{n}(\hat{\theta} - \theta)/\sigma \sim N(0, \Sigma). \tag{1.20}$$

Assuming $\lambda = \lambda_0$ is known, and let $\tilde{\beta} = \hat{\beta}(\lambda_0)$ is independent of $\tilde{\sigma} = \hat{\sigma}(\lambda_0)$. Therefore, it would be straightforward to make inference about parameters and it is given by

$$\sqrt{n}(\tilde{\beta} - \beta)/\sigma \sim N(0, Q^{-1}), \quad (n - p - 1)\tilde{\sigma}^2 \sim \chi^2_{n-p-1}, \tag{1.21}$$

and so we have

$$\frac{(\tilde{\beta} - \beta)^T X^T X(\tilde{\beta} - \beta)}{p\tilde{\sigma}^2} \sim F_{p,n-p-1}, \tag{1.22}$$

where $F_{p,n-p-1}$ and χ^2_{n-p-1} are F-distribution and χ^2 distribution respectively.

If we estimate λ from data, the variance of $\hat{\mu}$ and $\hat{\sigma}$ can obtained from eqn. (1.19) and (1.20). It presented that the variance of both parameters, $\hat{\mu}$ and $\hat{\sigma}$, can be increased as a result of λ estimation. There is the fact that using $F_{p,n-p-1}$ for the unconditional inference on β is not appropriate.

To construct a conditional inference in terms of λ, it was discussed by Chen and Lockhart [1997] in details. Here, denote parameter $\theta = (\sigma, \mu)$ and let $h = \sqrt{n}(\hat{\lambda} - \lambda)/\sigma$. Hence, we can compute the conditional distribution $\sqrt{n}(\hat{\theta} - \theta)/\sigma$ given $h = h_0$ in eqn. (1.20). It can be written,

$$\sqrt{n}(\hat{\theta} - \theta)/\sigma|_{h=h_0} \sim N(m_0, \Sigma_0), \tag{1.23}$$

and we have,

$$m_0 = \begin{bmatrix} a \\ Q^{-1}C \end{bmatrix} h_0, \tag{1.24}$$

and

$$\Sigma_0 = \begin{bmatrix} 1/2 & 0 \\ 0 & Q^{-1} \end{bmatrix}. \tag{1.25}$$

It seems that the covariance matrix for the conditional distribution is similar to the case λ known. In other words, F-distribution $F_{p,n-p-1}$ can be used to make the conditional inferences on β.

Chen et al. [2002] considered limits as $\delta \to 0$, likewise Bickel and Doksum [1981] also used a limit when $\delta \to 0$ as $n \to \infty$. They assumed that λ and β are fixed, however δ tends to zero as $\sigma \to 0$. Probability density function in Box-Cox model can be affected by parameters as assuming fixed n. Chen et al. [2002] defined the two parameters such as θ and ϕ as follows

$$\hat{\theta} = \hat{\beta}/\hat{\sigma}, \quad \hat{\phi} = \delta(\hat{\lambda} - \lambda)/\lambda. \tag{1.26}$$

Consequently, it was concentrated on the asymptotic expansions of $\hat{\phi}$ and $\hat{\theta}$ by considering a limit and specific conditions. Draper and Cox [1969] and Taylor [1986] also mentioned to employ small parameter δ for the same expansion.

In Chapter 2, our work would provide the parameter ξ which is related to parameter $\delta = \lambda\sigma/(1+\lambda\mu)$ indicated in Chen et al. [2002] and Bickel and Doksum [1981]. Bickel and Doksum provided a pretty poor approximation in asymptotic calculations of β around 0, this issue was criticized by several authors in application.

Yeo and Johnson [2000] assumed that transformed variables, $Y^{(\lambda)}(\lambda, Y_1)$, ..., $Y^{(\lambda)}(\lambda, Y_n)$ can be considered as a normal distribution for some λ. Therefore, log-likelihood is given by,

$$\log L(\theta, Y) = -n/2 \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(Y^{(\lambda)}(\lambda, y_i) - \mu\right)^2 + (\lambda - 1) \sum_{i=1}^{n} \mathrm{sgn}(\mathrm{y_i}) \log(|y_i| + 1), \tag{1.27}$$

where $\theta = (\sigma, \mu, \lambda)$ and $Y^{(\lambda)}(\lambda, y_i)$ normally distributed with μ and σ^2. Maximizing $L(\theta, Y)$ in terms of fixed λ, we have,

$$\hat{\mu}(\lambda) = 1/n \sum_{i=1}^{n} Y^{(\lambda)}(\lambda, y_i), \quad \hat{\sigma}^2(\lambda) = 1/n \sum_{i=1}^{n} \left(Y^{(\lambda)}(\lambda, y_i) - \hat{\mu}(\lambda)\right)^2. \tag{1.28}$$

Later, $\hat{\lambda}$ is computed by maximizing profile log-likelihood function and so we obtain $\hat{\theta} = (\hat{\sigma}^2(\hat{\lambda}), \hat{\mu}(\hat{\lambda}), \hat{\lambda})$.

McLeod [2009] discussed the best symmetrizing transformation for four different distributions in Mathematica demonstration. The probability density function for the transformed random variable computed where λ on [-2,2] and random variable with support on (0, ∞) in original domain. The specific value of λ was shown for each of the distribution that removes the skewness of transformed random variable for all the distribution used [McLeod, 2009]. In general, the value of λ may depend on the shape parameter in order to make symmetric distribution. McLeod [2009] stated that it would be possible to not find a symmetrizing transformation for some distributions including bimodal distribution. Box-Cox power transformations are presented to use normal curve theory for non normal distribution of random variable [Griffith, 2013].

1.3 Transformations and Unbounded Likelihood Problem

Atkinson and Pericchi [1991] discussed that ordinary maximum likelihood method behaves poorly for the random variable depending on an unknown parameter. Grouped likelihood approach was suggested by Atkinson and Pericchi [1991] for the shifted power transformation

[Box and Cox, 1964] to handle non-regular problems. It is assumed non-regular problem in the case that the distribution domain depends on the unknown shifted parameter. In theory, the maximum likelihood can not satisfy the regularity conditions if the range of the observations is defined by an unknown parameter. It is significant to denote the likelihood function as proportional of the probability functions. Montoya et al. [2009] criticized the strange behavior of profile likelihood functions which is derived by an unbounded density likelihood.

Cheng and Traylor [1995] expressed four types of non-regular problems in some situations and pointed out the unbounded likelihood as one of the special cases. Li et al. [2009] proposed EM algorithm for non-finite Fisher information if regularity conditions are failed to fulfill. In fact, unbounded behavior would lead to the problems in convergency and nonsense results for MLE. Liu et al. [2015] provided the "correct likelihood" to address the problem of unbounded likelihood by using small intervals.
Assuming a linear model for transformed values $Y(\lambda_1, \lambda_2)$ as follows,

$$Y(\lambda_1, \lambda_2) = X\beta + \epsilon. \tag{1.29}$$

Let $y_i^-(\lambda_1, \lambda_2)$ and $y_i^+(\lambda_1, \lambda_2)$ define transformation as $y_i + \Delta$ and $y_i - \Delta$ used in eqn. (1.4). So, the contribution of y_i in likelihood given by,

$$p_i = \frac{\Phi(w_i^+) - (w_i^-)}{1 - \Phi(-x_i\beta/\sigma)}, \tag{1.30}$$

where $w_i^\pm = (Y^\pm(\lambda_1, \lambda_2) - x_i\beta)/\sigma$ and Φ is standard normal distribution.
Grouped log-likelihood can be written as,

$$\log L(\lambda_1, \lambda_2, \beta, \sigma) = \sum_{i=1}^{n} \log p_i - n \log(2\Delta). \tag{1.31}$$

The correct likelihood was proposed by Liu et al. [2015] as a preliminary approach is given by,

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta; t_i) = \prod_{i=1}^{n} \frac{1}{\Delta_i} \left[F(t_i + \Delta_i; \theta) - F(t_i - \Delta_i; \theta)\right], \tag{1.32}$$

where θ is parameter and Δ defined as the round-off error. The round-off error would present the estimated error in calculation by using rounding.

The sensitivity analysis of Δ values was investigated by Atkinson and Pericchi [1991] and Liu et al. [2015]. It was argued about the effect of the round-off error on estimation of parameters precisely. Liu et al. [2015] also mentioned that the correct likelihood may not resolve unboundness of likelihood function in the case of multiple maximum or its flatness.

1.4 Non-Parametric Methods

Duan [1983] discussed the smearing estimate to predict the conditional mean of linear model after transformation. This model is non-parametric method used for expected response on the original domain. We have the estimation by the smearing estimate,

$$\hat{E}(Y_0) = \int h(x_0\hat{\beta} + \epsilon) d\hat{F}_n(\epsilon) = \frac{1}{n}\sum_{t=1}^{n} h(x_0\hat{\beta} + \hat{\epsilon}), \tag{1.33}$$

and, if the distribution of error F is not known, we estimate F function by empirical estimate as follows,

$$\hat{F}_n(e) = \frac{1}{n}\sum_{t=1}^{n} I(\hat{\epsilon}_i \leq e), \tag{1.34}$$

where $\hat{\epsilon}_i = Y_i^{(\lambda)} - x_i\hat{\beta}$ is the least squares residual and $I(.)$ is defined as an indicator function. We assume that g and h are monotone and continuous differentiable functions. Hence, it is defined by,

$$Y_i^{(\lambda)} = g(Y_i), \quad Y_i = h(Y_i^{(\lambda)}).$$

Consistency and efficiency of estimate were investigated by Duan [1983] and then compared with a parametric method in regression model. Taylor [1986] also compared the conditional mean by smearing estimate and Taylor expansion in linear model. Gibbs Sampling can compute any desired expectation from posterior distribution. This approach is one of MCMC technique. It was argued by Taylor [1986] that the bias of the small-θ approximation method could be reduced by using the higher order of expansion for the conditional mean, while the smearing estimator would tend to decrease variation of variance.

Breiman and Friedman [1985] provided non-parametric method to find optimal transformation in multiple regression and stationary time series. The aim of this approach is the same as the Box-Cox transformation method. Alternating conditional expectation (ACE) algorithm was applied to different dataset for comparison. Let X_1, ..., X_p be mean zero stationary time series and θ, ϕ_1, ..., ϕ_p are real valued function. Thus, the optimal transformations are computed by minimizing the following function,

$$e^2 = \frac{E\left[\theta(X_{p+1}) - \sum_{i=1}^{p}\phi_i(X_i)\right]^2}{E[\theta^2(X_{p+1})]}. \tag{1.35}$$

The procedure was initially implemented by Breiman and Friedman [1985] for simulated data when optimal transformations are given, and then it was applied to the Boston housing data of Harrison and Rubinfeld [1978].

A non-parametric estimation method was suggested by Han [1987] for the transformed model using Kendall's rank correlation, and it was discussed to be more consistent and efficient than the maximum likelihood estimator. A simple semi-parametric estimation method

is introduced by Foster et al. [2001] for the Box-Cox transformation without assuming normal distribution of the error term . It was illustrated by numerical simulation for the specific dataset, and also it was derived that estimators are consistent and asymptotically normal [Foster et al., 2001].

1.5 Box-Cox Transformations and Time Series

Maximum likelihood method applied to estimate the Box-Cox transformation parameter λ and confidence interval for λ in Box-Cox transformed family in AR model as follows,

$$z_t^{(\lambda)} = \begin{cases} (z_t^{\lambda} - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log(z_t), & \text{if } \lambda = 0. \end{cases} \tag{1.36}$$

for time series data z_t, $t = 1, ..., n$. Hipel and McLeod [1994] discussed this method for seasonal and non-seasonal ARIMA model to obtain the $z_t^{(\lambda)}$ series. Finally, the back-transformed time series would be directly calculated from the ARMA or ARIMA model in the original data domain. Let log-likelihood function for an assumed value λ in AR(p) model be defined by,

$$\log L(\phi, \lambda) = -\frac{n}{2}\log(S(\phi)/n) - \frac{1}{2}\log(g_n) + (1-\lambda)\sum_{t=1}^{n}\log(z_t), \tag{1.37}$$

and then maximizing over ϕ leads to $L(\lambda)$ for λ. Using optimize function to maximize $L(\lambda)$ function numerically which $\hat{\lambda}$ obtained. The relative likelihood function plot, $R(\lambda) = L(\lambda)/L(\hat{\lambda})$, illustrated a 95% confidence interval for λ. Box et al. [2008] discussed that the use of the Box-Cox transformation may improve the accuracy of the forecasts. Later, the Box-Cox transformation considered by Proietti and Riani [2009] for positive time series and multivariate time series. By using numerical and Monte Carlo integration, two conditional moments of seasonally adjusted time series were computed [Proietti and Riani, 2009]. Proietti and Riani [2009] developed a Taylor series expansion to determine the inverse transformation.
The optimal forecast for the seasonally adjusted series can be written as,

$$\hat{z}_t = E(z_t|F_t) = \int_{-\infty}^{+\infty} (\lambda z_t^{(\lambda)} + 1)^{1/\lambda} f(z_t^{(\lambda)}|F_t) dz_t^{(\lambda)}, \tag{1.38}$$

and, the conditional variance of the forecast error is presented by,

$$Var(z_t|F_t) = \int_{-\infty}^{+\infty} (z_t - \hat{z}_t)^2 f(z_t^{(\lambda)}|F_t) dz_t^{(\lambda)}. \tag{1.39}$$

The naive forecasts can be simply obtained by,

$$\hat{z}_t = \begin{cases} (1 + \lambda \hat{z}_t^{(\lambda)})^{1/\lambda}, & \text{if } \lambda \neq 0, \\ \exp(\hat{z}_t^{(\lambda)}), & \text{if } \lambda = 0. \end{cases} \tag{1.40}$$

Granger and Newbold [1976] employed the Hermit polynomial expansion to compare the autocorrelation of data in the original data domain and transformed data domain. This method can be used by considering the Gaussian assumption of transformed time series. Consequently, they expressed that the original series is always less forecastable compare to the transformed series. Forecastability can be presented by [Granger and Newbold, 1976],

$$R^2_{h,z_t} < R^2_{h,z_t^{(\lambda)}}, \tag{1.41}$$

where h is a lead time. Further, the autocorrelation of the transformed stationary Guassian time series can be defined as $corr(z_t^{(\lambda)}, z_{t-k}^{(\lambda)}) = \rho_{z_t^{(\lambda)}}(k)$. It was shown that,

$$|\rho_{z_t}(k)| < |\rho_{z_t^{(\lambda)}}(k)|. \tag{1.42}$$

Furthermore, Granger and Newbold [1976] investigated loss function in the case of mean square error and mean absolute error. It needs to develop a numerical method for obtaining the optimal forecast for any specified loss function. The fact that the expected value of the inverse Box-Cox transformed can be considered as the minimum mean square error (MMSE) prediction, and the variance of the inverse transformed is its mean square error (MSE). Granger and Newbold [1976] discussed how their method can be extended to the homogenous non-stationary time series models which correspond to spatial models with the intrinsic stationary assumption.

1.6 Illustrative Application

It has been a problem to predict the sunspot numbers time series for several researchers and different forecasting methods have carried out in order to obtain the optimal forecast. In this section, the main goal is to explore the effect of two power transformations on the fitting model and forecasting.

To illustrate the discrepancy between the Yeo-Johnson transformation and the Box-Cox transformation, sunspot time series dataset would be considered. Firstly, we apply both transformations discussed on monthly sunspot numbers from 1749 to 1983 and also yearly average sunspot numbers from 1700 to 1988 to obtain the optimal transformation. These two time series investigated in this study both include zero values, even though Box and Cox [1964] defined the Box-Cox transformations only for positive random variables.

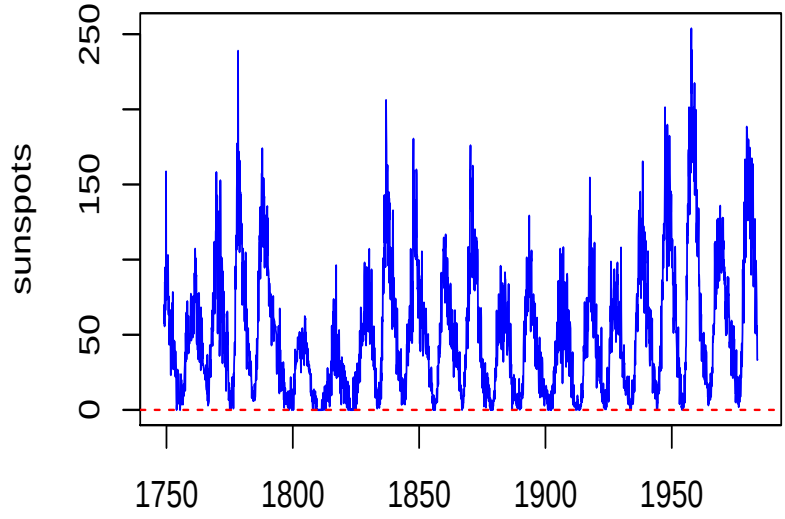

Figure 1.1: Time series plot of sunspots numbers.

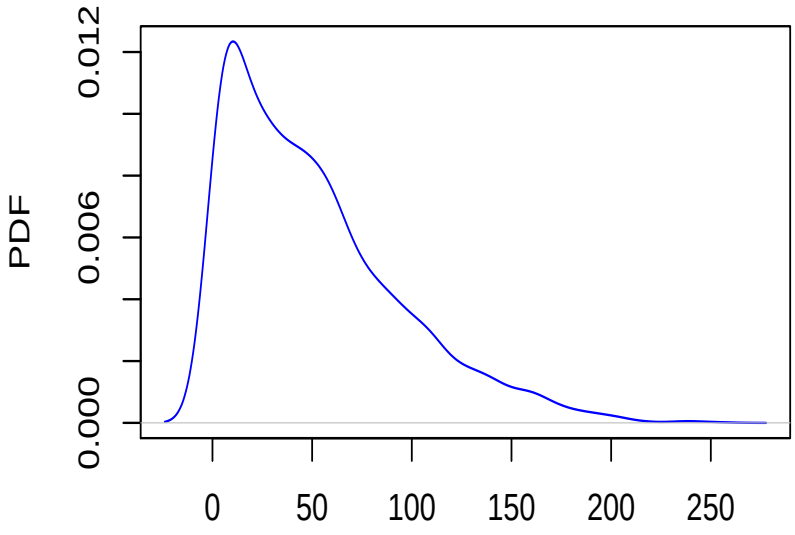

Figure 1.2: The probability density function using Gaussian kernel of sunspots.

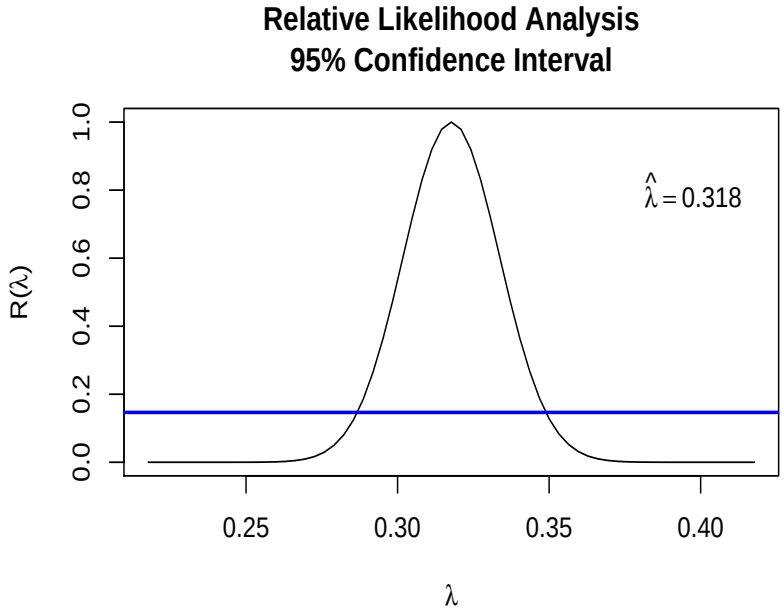

Figure 1.3: Yeo-Johnson transformations was used.

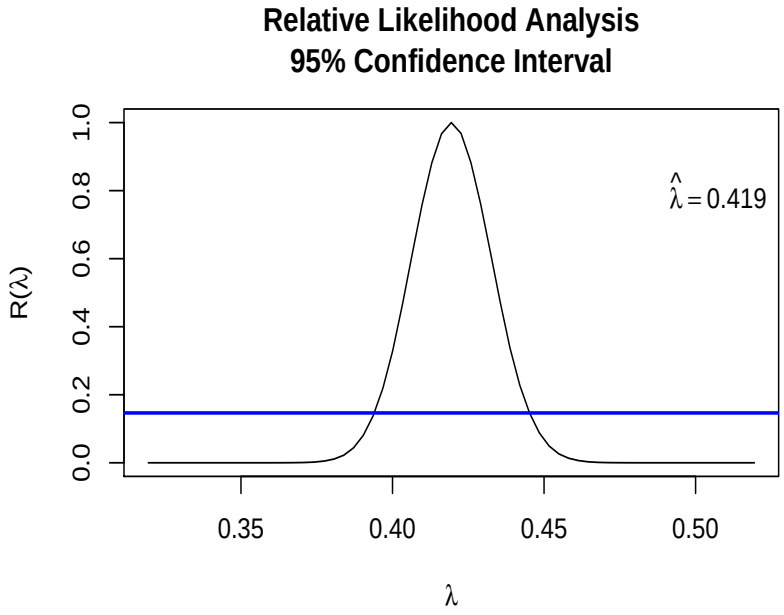

Figure 1.4: Box-Cox transformations was used.

There is no restriction on the transformation defined by Yeo and Johnson [2000]. This transformation like the Box-Cox transformation is invertiable, and also the inverse transformed

variables can be derived. There is the fact that the Yeo-Johnson transformation can cover all range of $(-\infty, \infty)$, hence it maybe provide more exact analysis compared to the Box-Cox transformation.

Therefore, we apply relative likelihood function to determine the optimal transformation, $\hat{\lambda}$. From Figure 1.3 and 1.4, we can conclude that $\hat{\lambda}$ based on the Yeo-Johnson transformation is slightly smaller compared to the Box-Cox transformation for monthly sunspots time series. Furthermore, the Yoe-Johnson transformation produces slightly wider confidence interval for λ when $Y > 0$, and also $\hat{\lambda} = 0.419$ is not included in the confidence interval of the Box-Cox method. To produce Figure 1.3 and 1.4, additive shift used in both transformation. The 95% confidence interval for λ can be obtained by $\log L(\hat{\lambda}) - \log L(\lambda) < 1/2\chi^2_{1,(1-\alpha)}$.

The boxplot shown below reveals that the Yeo-Johnson transformed data is more variable, but there is still some evidence of left skewness. The skewnesses are -0.14 and -0.20 for the Box-Cox and Yeo-Johnson transformed data respectively.

Figure 1.6 illustrates the normal probability of the residuals. The Box-Ljung portmanteau diagnostic plots produced for the Yeo-Johnson and Box-Cox transformed time series are shown in Figures 1.7 and 1.8. These diagnostic plots confirm that the AR(9) is a reasonable model for the both of the transformed data. From Figure 1.6, after transformation of the original time series, we can have an error distribution which behaves normal.

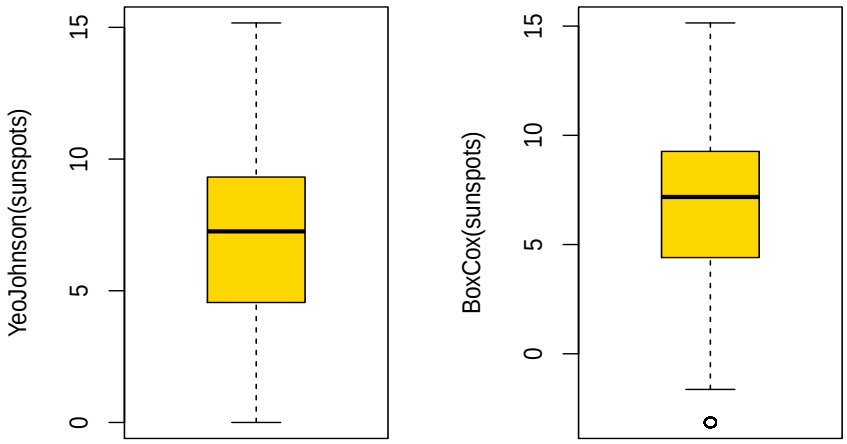

Figure 1.5: Comparison of the Yeo-Johnson transformed and Box-Cox transformed of sunspots data set.

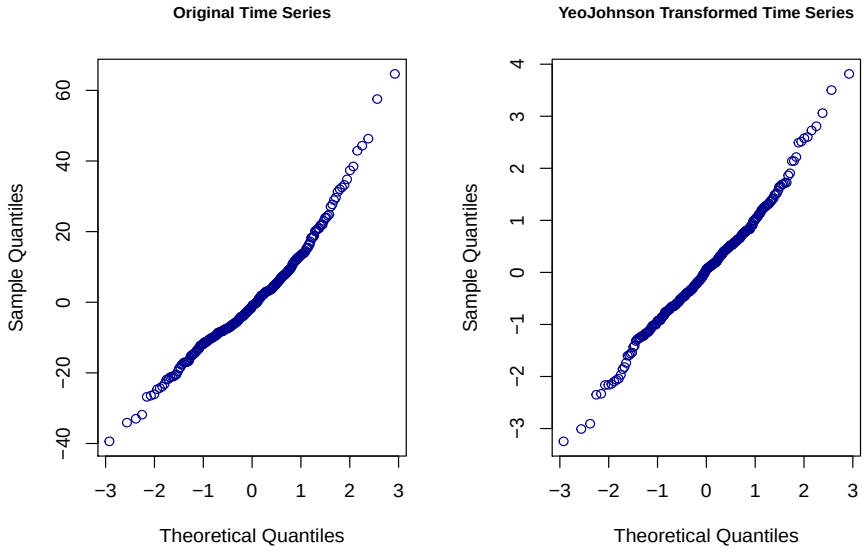

Figure 1.6: Normal probability plot of the standardized prediction residuals of the fitted AR(9) model to original time series and Yeo-Johnson transformed time series with $\lambda = 0.318$.

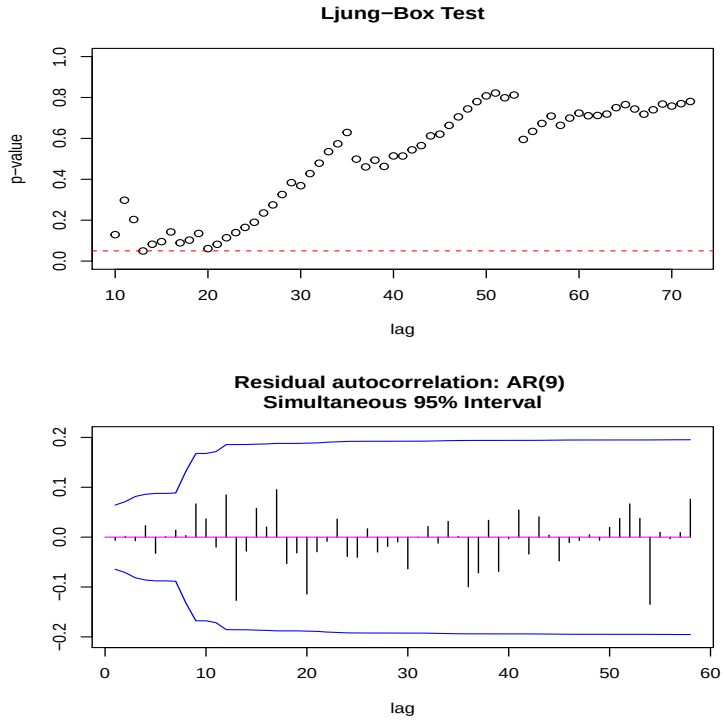

Figure 1.7: Diagnostic plots produced for AR(9) model fit to the Yeo-Johnson transformation of yearly sunspot series.

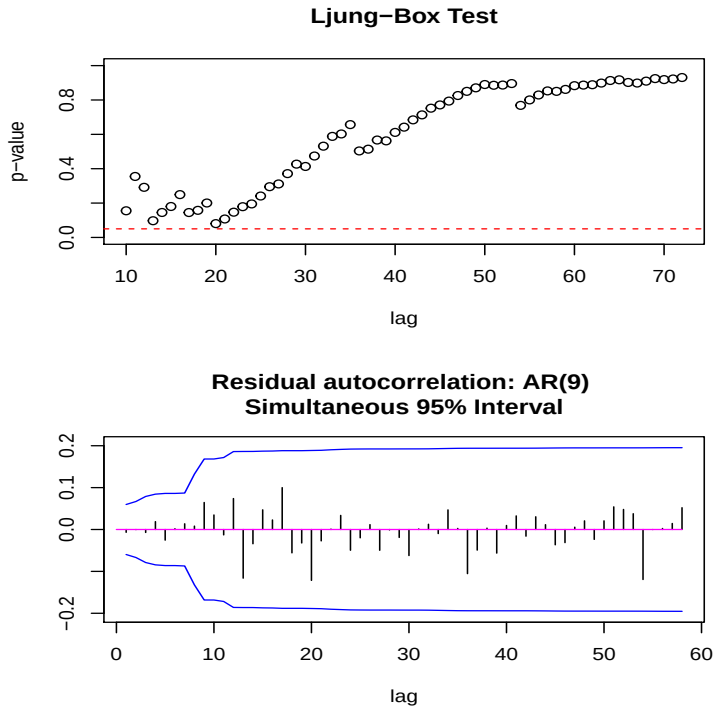

Figure 1.8: Diagnostic plots produced for AR(9) model fit to the Box-Cox transformation of yearly sunspot series.

	BoxCox(sunspot.year)		YeoJohnson(sunspot.year)	
Model	AIC	Portmanteau Diagnostic	AIC	Portmanteau Diagnostic
AR(2)	346.8	fail	120.6	fail
AR(9)	288.2	satisfactory	59.3	satisfactory
ARMA(2,1)	1168.62	borderline	942	borderline

Table 1.1: Models fit to transformed yearly sunspot numbers time series.

Several models were evaluated to obtain the long-term prediction of the transformed sunspot numbers, and also their performance is summarized in Table 1.1. Now, we test the performance of the Box-Cox and Yeo-Johnson transformation on forecasts of the time series when it was fitted to AR(9) model. In Table 1.2, we compare the forecasts at origin time $n = 289$ for lead times $l = 1, 2, 3$ for the AR(9) model. In addition, the standard deviations of the forecasts were computed and its values are significantly higher for approximate Box-Cox transformation. The differences between the forecasts from the Box-Cox and Yeo-Johnson transformations are substantial in the transformed scale.

	BoxCox(sunspot.year)		YeoJohnson(sunspot.year)	
Lead	Forecast	Standard deviation of forecast	Forecast	Standard deviation of forecast
1	17.271	1.583	12.419	1.065
2	17.877	2.483	12.728	1.662
3	17.129	2.925	12.283	1.948

Table 1.2: Forecasts and their standard deviations for fitted AR(9) model to transformed sunspot.year time series in terms of the Box-Cox and Yeo-Johnson transformations.

1.7 Maximum Likelihood Estimation

Fisher [1922] introduced the maximum-likelihood estimation technique following by Wald [1949] that discussed the asymptotic properties of MLE. The idea of modified likelihood method developed when the complete likelihood is difficult or impossible to calculate. Cox [1975] proposed the partial likelihood in the case it is more simpler than complete likelihood and also it only contains parameter interest rather than nuisance parameters. Conditional and marginal Likelihood methods are proposed to deal with some multiparameter problems. Statistical inferences about the parameters of interest can be determined by eliminating nuisance parameters from likelihood function [Kalbfleisch and Sprott, 1970].

Assume X_1, ..., X_n be a random variables from a population whose density depends on the parameters θ and δ where are called a structural and incidental parameters respectively. To obtained the estimate of θ, conditional distribution was considered given minimal sufficient statistics for δ. Define $T_i = T(x_i)$ be the minimal sufficient statistic for δ_i and also probability distributions of T_i be called $g(t_i|\theta, \delta)$. Thus, conditional distribution of X_1, X_2,..., X_n given $T_1 = t_1$,..., $T_n = t_n$ can be written [Andersen, 1970] as,

$$\phi(x_1, ..., x_n|\theta, t_1, ..., t_n) = \prod_{i=1}^{\infty} \phi(x_i|\theta, t_i) = \prod_{i=1}^{\infty} f(X_i|\theta, \delta_i)/g(t_i|\theta, \delta), \tag{1.43}$$

where T_i is sufficient statistics for δ due to independency of δ. The asymptotic normality of the conditional maximum-likelihood estimation under assumptions defined as follows.

Theorem 1.7.1 *The first and second derivatives of* $\log f(x_i, \theta, t)$ *with respect to* θ *exist for all* θ *in an open interval* Θ *, and for all* δ *is given by,*

$$E\left(\partial \log \phi(x_i|\theta, T)/\partial\theta\right) = 0, \tag{1.44}$$

and $E\left[\partial^2 \log \phi(x_i|\theta, T)/\partial\theta^2\right] > 0$ and be continuous function of δ.

Assuming that the model is correct, the likelihood principle, the likelihood function contains all the information needed for statistical inference on the parameters. Kalbfleisch and

Sprott [1970, 1973] described how conditional likelihoods may be useful in eliminating nuisance parameters when the likelihood can be factored into two parts. There are two aspects to consider here would not be straightforward to express mathematically. First, variables X should include the all information required for the parameters of interest. Further, the distribution of X is dependent on nuisance parameters. Secondly, nuisance parameters should not appear in the partial likelihood.

1.8 EM Algorithm

Statistical inference was mostly determined by MLE method as a result of its asymptotic normality and efficiency properties. We propose EM algorithm which is more flexible and reliable to estimate parameters for truncated data. The EM algorithm is preferable over the numerical optimization because at each iteration the likelihood function increases and also the rate of convergency implies to stationary point. The EM process can be employed to determine the maximum likelihood estimate for censored and truncated data which come from exponential family [Dempster et al., 1977]. Lee and Scott [2012] illustrated the EM algorithm to fit multivariate Gaussian mixture models on truncated and censored data. In general, if $L(\theta|y)$ has several stationary points, the convergency of EM sequence to local or global maximizers and saddle points depends on the choice of initial point θ_0 [Wu, 1983]. Cauchy distribution can be considered as a non-regular case which its likelihood function with respect to location parameter can be multimodal. Simulated annealing technique is performed to reach a global MLE with high probability for this special situations [Robert and Casella, 2004]. Furthermore, genetic optimization or Monte-Carlo Markov Chain (MCMC) would be applicable for this type of problem.

The EM procedure is very popular for computing maximum likelihood estimates from incomplete data, despite that fact that numerical optimization may be converged slowly. In the case where the likelihood function satisfies regularity conditions and $L(\theta|y)$ is unimodal undefined domain, the EM process convergences to unique MLE [McLachlan and Krishnan, 2007]. Dempster et al. [1977] suggested an algorithm to compute iteratively maximum likelihood estimates for incomplete data including censored and truncated data. The EM approach contains two steps which each iteration of the expectation step (E-step) followed by the maximization step (M-step).
Suppose that observed data Y_i, $i = 1, ..., n$ have probability density function $g(y|\theta)$. Then we can write,

$$g(y|\theta) = \int_Z f(y, z|\theta)dz, \tag{1.45}$$

and our main purpose is that the parameter can be obtained by,

$$\hat{\theta} = \text{argmax}\, L(\theta|y) = \text{argmax}\, g(y|\theta). \tag{1.46}$$

The log-likelihood for observed data, Y, is given by,

$$\log L(\theta|y) = \log(g(y|\theta)). \tag{1.47}$$

We assume that (Y, Z) as complete data have PDF $f(y, z, \theta)$ and log-likelihood of complete data can be written,

$$\log L^c(\theta|z, y) = \log(f(y, z|\theta)). \tag{1.48}$$

The conditional density of incomplete data Z given observed data Y and θ then becomes,

$$k(z|y, \theta) = \frac{f(y, z|\theta)}{g(y|\theta)}. \tag{1.49}$$

So that, by taking logs

$$\log g(y|\theta) = \log f(y, z|\theta) - \log k(z|y, \theta). \tag{1.50}$$

We can define for given θ_0,

$$E_{\theta_0} \log L(\theta|y) = E_{\theta_0}[\log L^c(\theta|z, y)] - E_{\theta_0}[\log k(z|y, \theta)], \tag{1.51}$$

where the expectation define in terms of distribution $k(z|y, \theta_0)$. We only consider the first term on the right side of eqn. (1.51) to achieve a maximizing $\log L(\theta|y)$. By assuming the interchange expectation with respect to Z and differentiation in terms of θ_0, let us have

$$\partial_{\theta_0} E_{\theta_0}[\log k(z|y, \theta)] = E_{\theta_0} \partial_{\theta_0}[\log k(z|y, \theta)] = 0. \tag{1.52}$$

We consider the theory that the expectation of score function become zero [Casella and Berger, 2002]. We can conclude that $\partial_{\theta_0} E_{\theta_0}[\log k(z|y, \theta)]$ is not depend on parameter θ, and then we can maximize $E_{\theta_0}[\log L^c(\theta|z, y)]$.

We aim to maximize $Q(\theta|\theta_0, y)$. Let us define,

$$Q(\theta|\theta_0, y) = E_{\theta_0}[\log L^c(\theta|z, y)]. \tag{1.53}$$

The iterative process begin with a given initial value θ_0 and let θ_j denote the value of θ after j cycles. The next cycle can be processed in two steps as follows,

1. E-step: compute the expected log-likelihood function of complete data,

$$Q(\theta|\hat{\theta}_j, y) = E_{\hat{\theta}_j}[\log L^c(\theta|z, y)]. \tag{1.54}$$

2. M-step: determine the parameter θ_{j+1} that maximize likelihood,

$$\hat{\theta}_{j+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\hat{\theta}_j, y). \tag{1.55}$$

1.8.1 General Properties of EM Algorithm

The EM process can be applied when data come from exponential family and then it is solvable under the convexity property of log-likelihood including JensenâĂŹs inequality and the Kullback-Liebler discrepancy. We would present these concepts more in details and then use them in the derivation of the EM algorithm.

Theorem 1.8.1 *Define $f : X \to R$ be a convex function if $\forall x_1, x_2 \in \mathbb{X}$, and $\forall t \in [0, 1]$ we have,*

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2), \tag{1.56}$$

so, it is called strictly convex if equality holds for $t = 0$ or $t = 1$.

Theorem 1.8.2 *Let p_1, ..., p_n be $Pr(X_i = x) = p_i$ and f is a real continuous function which is convex. Then Jensen's inequality given by,*

$$f(\sum_{i=1}^{n} p_i x_i) \leq \sum_{i=1}^{n} p_i f(x_i). \tag{1.57}$$

In general, we assume x as a random variable and f as any convex function where E denoted expectation. Hence, it follows from Jensen's inequality,

$$f(E(x)) \leq E(f(x)). \tag{1.58}$$

Theorem 1.8.3 *Jensen's inequality is used to create the non-negativity of the Kullbach-Liebler discrepancy. Denote f(x) and g(x) be two probability density functions on R, the Kullbach-Liebler discrepancy is given by,*

$$K(g, f) = \int \log \frac{f(x)}{g(x)} f(x)dx = E_f \log \frac{f(x)}{g(x)}. \tag{1.59}$$

Proof

$$K(g, f) = \int \log (\frac{f(x)}{g(x)}) f(x)dx = -\int \log (\frac{g(x)}{f(x)}) f(x)dx \geq -\log \int (\frac{g(x)}{f(x)}) f(x)dx = 0 \tag{1.60}$$

We apply all theorem defined to show that likelihood increases at each step of EM algorithm. Using the sequences of $\hat{\theta}_j$, $j = 0, 1, 2, ...$ given by the EM algorithm satisfy,

$$\log L(\hat{\theta}_{j+1}|y) \geq \log L(\hat{\theta}_j|y). \tag{1.61}$$

Equality yields in eqn. (1.61) if and only if,

$$Q(\hat{\theta}_{j+1}|\hat{\theta}_j, y) = Q(\hat{\theta}_j|\hat{\theta}_j, y), \tag{1.62}$$

Proof By using eqn. (1.51), it can be written as,

$$\log L(\theta|y) = Q(\theta|\hat{\theta}_j, y) - E_{\theta_j} \log k(z|y, \theta_j). \tag{1.63}$$

Hence we have,

$$\log L(\hat{\theta}_{j+1}|y) = Q(\hat{\theta}_{j+1}|\hat{\theta}_j, y) - E_{\theta_j} \log k(z|y, \theta_{j+1}), \tag{1.64}$$

and

$$\log L(\hat{\theta}_j|y) = Q(\hat{\theta}_j|\hat{\theta}_j, y) - E_{\theta_j} \log k(z|y, \theta_j). \tag{1.65}$$

Therefore, it was gained

$$\log L(\hat{\theta}_{j+1}|y) - \log L(\hat{\theta}_j|y) = Q(\hat{\theta}_{j+1}|\hat{\theta}_j, y) - Q(\hat{\theta}_j|\hat{\theta}_j, y) - E_{\theta_j} \log k(z|y, \theta_{j+1}) + E_{\theta_j} \log k(z|y, \theta_j). \tag{1.66}$$

We present that by using $Q(\hat{\theta}_{j+1}|\hat{\theta}_j, y) - Q(\hat{\theta}_j|\hat{\theta}_j, y) \geq 0$, eqn. (1.61) holds if we have,

$$E_{\theta_j} \log k(z|y, \theta_{j+1}) \leq E_{\theta_j} \log k(z|y, \theta_j). \tag{1.67}$$

This can be written as,

$$E_{\theta_j} \log \frac{k(z|y, \theta_{j+1})}{k(z|y, \theta_j)} \geq 0. \tag{1.68}$$

It follows from Jensen's inequality and Kullback-Liebler discrepancy definition.

1.9 Appendix. Information Matrix

Under regularity conditions, the MLE of $\hat{\theta}$ is a consistent estimator of θ and also the asymptotic distribution of $\sqrt{n}(\hat{\theta}-\theta)$ is $N(0,\Sigma)$ where Σ can be consistently estimated by $\hat{\Sigma} = -\left[\partial^2 L/\partial\theta\partial\acute{\theta}\right]$ assessed at $\theta = \hat{\theta}$ [Cox and Hinkley, 1979]. Therefore, the information matrix is appropriate approach to estimate the approximate standard errors of the MLE estimates. For comparison with the truncated case, we first provide the result for random sampling from a complete normal distribution. It would be simpler to consider σ rather than σ^2 in finding the information matrix . In the normal IID case with complete data for a random sample of size n from a normal population with mean μ and variance σ^2, the Fisher information matrix for (μ,σ) can be defined by,

$$I(\mu,\sigma) = \begin{bmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{2n}{\sigma^2} \end{bmatrix}. \tag{1.69}$$

In this part, we calculated expected information matrix in the more general case with considering the truncation effect. Hence, denote

$$\sum_{i=1}^{n} I_i(\sigma,\mu,\beta,\lambda) = \begin{bmatrix} i_{11} & i_{12} & i_{13} & i_{14} \\ i_{21} & i_{22} & i_{23} & i_{24} \\ i_{31} & i_{32} & i_{33} & i_{34} \\ i_{41} & i_{42} & i_{43} & i_{44} \end{bmatrix}. \tag{1.70}$$

The log-likelihood, $l(\mu,\sigma,\lambda)$, can be written as follows,

$$\log L(\mu,\sigma,\lambda) = \sum_{i=1}^{n} \log\left(\phi\left(\frac{Y_i^{(\lambda)}-\mu_i}{\sigma}\right)\right) + (\lambda-1)\sum_{i=1}^{n}\log\left(Y_i\right) - n\log(1-\Phi(\xi)). \tag{1.71}$$

Taking the first derivatives,

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n}(Y_i^{(\lambda)}-\mu_i)/\sigma^2 - n(1/\sigma)\frac{\phi(\xi)}{1-\Phi(\xi)}, \tag{1.72}$$

where $\Psi(\xi) = \phi(\xi)/1-\Phi(\xi)$ and $\xi = (T-\mu_i)/\sigma$.
From eqn. (1.72), we can derive i_{22},

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} - (\frac{n}{\sigma})\left(\frac{\partial \phi_\mu(\xi)(1-\Phi(\xi)) - (1-\partial\Phi_\mu(\xi))\phi(\xi)}{(1-\Phi(\xi))^2}\right)$$

$$= -\frac{n}{\sigma^2} - (\frac{n}{\sigma^2})\left(\frac{\xi\phi(\xi)(1-\Phi(\xi)) - \phi^2(\xi)}{(1-\Phi(\xi))^2}\right) \tag{1.73}$$

$$= -n\sigma^{-2} - n\sigma^{-2}\left[\xi\Psi(\xi) - \Psi^2(\xi)\right].$$

So, we have

$$i_{22} = -E\left[\frac{\partial^2 l}{\partial \mu^2}\right] = n\sigma^{-2}\left(1 + \xi E[\Psi(\xi)] - E[\Psi^2(\xi)]\right), \tag{1.74}$$

and also, we can derive i_{13} from eqn. (1.72)

$$\frac{\partial^2 l}{\partial \mu \partial \lambda} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \dot{Y}_i^{(\lambda)}. \tag{1.75}$$

Then, we obtain

$$i_{24} = -E\left[\frac{\partial^2 l}{\partial \mu \partial \lambda}\right] = -\frac{1}{\sigma^2}\sum_{i=1}^{n} E[\dot{Y}_i^{(\lambda)}]. \tag{1.76}$$

To obtain the first derivative with respect to λ,

$$\frac{\partial l}{\partial \lambda} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(Y_i^{(\lambda)} - \mu_i\right)\dot{Y}_i^{(\lambda)} + \sum_{i=1}^{n}\log Y_i. \tag{1.77}$$

Next, differentiating with respect to λ

$$\frac{\partial^2 l}{\partial \lambda \partial \lambda} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\ddot{Y}_i^{(\lambda)}(Y_i^{(\lambda)} - \mu_i) + (\dot{Y}_i^{(\lambda)})^2\right]. \tag{1.78}$$

Hence, we derive i_{33}

$$i_{44} = -E\left[\frac{\partial^2 l}{\partial \lambda \partial \lambda}\right] = -\frac{1}{\sigma^2}\sum_{i=1}^{n} E\left[\ddot{Y}_i^{(\lambda)}(Y_i^{(\lambda)} - \mu_i) + \dot{Y}_i^{(\lambda)}\right]. \tag{1.79}$$

Taking the second partial derivative with respect to σ from eqn. (1.72),

$$\frac{\partial^2 l}{\partial\mu\partial\sigma} == -2\sigma^{-3}\sum_{i=1}^{n}(Y_i^{(\lambda)} - \mu_i) + n\sigma^{-2}\left[\Psi(\xi) + \xi(\xi\Psi(\xi) + \Psi^2(\xi))\right], \tag{1.80}$$

so, we have

$$i_{21} = -E\left[\frac{\partial^2 l}{\partial\mu\partial\sigma}\right] = -n\sigma^{-2}\left[E[\Psi(\xi)] + \xi(\xi E[\Psi(\xi)] + E[\Psi^2(\xi)])\right]. \tag{1.81}$$

To derive i_{11}, we obtain the second partial derivative with respect to σ,

$$\frac{\partial^2 l}{\partial\sigma\partial\sigma} = n\sigma^{-2} - 3\sigma^{-4}\sum_{i=1}^{n}(Y_i^{(\lambda)} - \mu_i)^2 - n\sigma^{-2}\xi\left[-2\Psi(\xi) + \xi(\xi\Psi(\xi) - \Psi^2(\xi))\right]. \tag{1.82}$$

Thus, it given by

$$i_{11} = -E\left[\frac{\partial^2 l}{\partial\sigma\partial\sigma}\right] = 2n\sigma^{-2} + n\sigma^{-2}\xi\left[-2E[\Psi(\xi)] + \xi(\xi E[\Psi(\xi)] - E[\Psi^2(\xi)])\right]. \tag{1.83}$$

From eqn. (1.77), we can obtain

$$\frac{\partial^2 l}{\partial\lambda\partial\sigma} = 2\sigma^{-3}\sum_{i=1}^{n}\dot{Y}_i^{(\lambda)}(Y_i^{(\lambda)} - \mu_i). \tag{1.84}$$

Then, we have

$$i_{14} = -E\left[\frac{\partial^2 l}{\partial\lambda\partial\sigma}\right] = -2\sigma^{-2}\sum_{i=1}^{n}E[\dot{Y}_i^{(\lambda)}(Y_i^{(\lambda)} - \mu_i)/\sigma], \tag{1.85}$$

and

$$i_{34} = -E\left[\frac{\partial^2 l}{\partial\lambda\partial\beta}\right] = -\sigma^{-2}\sum_{i=1}^{n}E[\dot{Y}_i^{(\lambda)}x_i], \tag{1.86}$$

and

$$i_{33} = -E\left[\frac{\partial^2 l}{\partial\beta\partial\beta}\right] = \sigma^{-2}\sum_{i=1}^{n} E[XX^T]. \tag{1.87}$$

The average Fisher information matrix $I = \sigma^2/n\left(\sum_{i=1}^{n} I_i(\sigma,\mu,\beta,\lambda)\right)$ can be written by,

$$\frac{1}{n}\sum_{i=1}^{n}\sigma^2 I_i = \begin{bmatrix} 2+\xi w[-2+\xi^2-\xi w] & -w[1+\xi^2+\xi w] & 0 & -2a \\ -w[1+\xi^2+\xi w] & 1+w(\xi-w) & 0 & -b \\ 0 & 0 & Q & -C \\ -2a & -b & -C^T & d \end{bmatrix} \tag{1.88}$$

where

$$w = E[\Psi(\xi)], \tag{1.89}$$

$$a = 1/n\sum_{i=1}^{n} E\left[\frac{\dot{Y}_i^{(\lambda)}(Y_i^{(\lambda)}-\mu_i)}{\sigma}\right], \tag{1.90}$$

$$b = 1/n\sum_{i=1}^{n} E\left[\dot{Y}_i^{(\lambda)}\right], \tag{1.91}$$

$$C = 1/n\sum_{i=1}^{n} E\left[\dot{Y}_i^{(\lambda)}x_i\right], \tag{1.92}$$

$$d = 1/n\sum_{i=1}^{n} E\left[(Y_i^{(\lambda)}-\mu_i)\ddot{Y}_i^{(\lambda)} + (\dot{Y}_i^{(\lambda)})^2\right], \tag{1.93}$$

$$Q = 1/nX^TX, \tag{1.94}$$

where $\Psi(.) = \phi(.)/(1-\Phi(.))$ as $\lambda > 0$. In addition, $Y_i^{(\lambda)}$ assumed as the Box-Cox transformed variables where $\dot{Y}_i^{(\lambda)}$ and $\ddot{Y}_i^{(\lambda)}$ are the first and second derivatives of $Y_i^{(\lambda)}$. Therefore, the Fisher information for the exact Box-Cox likelihood can be defined using eqn. (1.88). By obtaining the Fisher information and its inverse, it can be noticed that the variance of the parameter estimators are dependent on $\Psi(\xi)$ when $\lambda > 0$. In other words, the first-order and second-order derivatives of log-likelihood in eqn. (1.71) associated with computing repeatedly these two fractions as follows. For $\lambda > 0$,

$$\Psi(\xi) = \phi(\xi)/1-\Phi(\xi), \tag{1.95}$$

and for $\lambda < 0$,

$$\Psi(\xi) = \phi(\xi)/\Phi(\xi), \tag{1.96}$$

where these fractions are the inverse Mills ratio, and also eqn. (1.95) denoted as hazard rate. Johnson and Kotz [1970] suggested the procedure to determine approximation for the Mills ratio.

Chapter 2

Exact Box-Cox Analysis

We will be working with various normal distributions so we introduce some convenient notations. Let $\phi(z, \mu, \sigma)$ and $\Phi(z, \mu, \sigma)$ denote the probability density function (PDF) and the cumulative density function (CDF) of a normal distribution with mean μ and standard deviation σ and let $\phi(z) = \phi(z, 0, 1)$ and $\Phi(z) = \Phi(z, 0, 1)$. The left and right truncated normal distribution functions with truncation point T are denoted respectively by $\phi_{(T,\infty)}(z, \mu, \sigma)$ and $\Phi_{(-\infty,T)}(z, \mu, \sigma)$.

2.1 Introduction

Box and Cox [1964] introduced the idea for selecting a suitable power transformation by considering the transformation as part of an enlarged model and showing how the transformation may be estimated by maximum likelihood (MLE). The Box-Cox transformation for random variable Y is defined for $Y > 0$ by

$$Z = Y^{(\lambda)} = \begin{cases} \left(Y^{\lambda} - 1\right)/\lambda, & \lambda \neq 0, \\ \log(Y), & \lambda = 0. \end{cases} \tag{2.1}$$

With this definition the transformation is a continuous function of λ so the Jacobian of the transformation $Y \to Z$ is a continuous and one-to-one.

Let Y be a positive random variable with probability density function (PDF) $f_Y(y)$. Then the PDF for $Z = Y^{(\lambda)}$ may be written,

$$f_Z(z) = \begin{cases} f_Y((1 + z\lambda)^{1/\lambda})(1 + z\lambda)^{1/\lambda - 1}, & \lambda \neq 0, \quad \lambda z + 1 > 0, \\ e^z f_Y(e^z), & \lambda = 0. \end{cases} \tag{2.2}$$

The transformation is used to improve the accuracy of the assumption of the normal distribution when the data exhibit skewness and other related non-normal features such outliers and monotone variance change. Often the transformation also improves the additivity assumption when used with the normal linear model. Figure 2.1 shows how the Weibull distribution can be made approximately normal with a suitable choice of λ, viz. $\lambda = 0.416$.

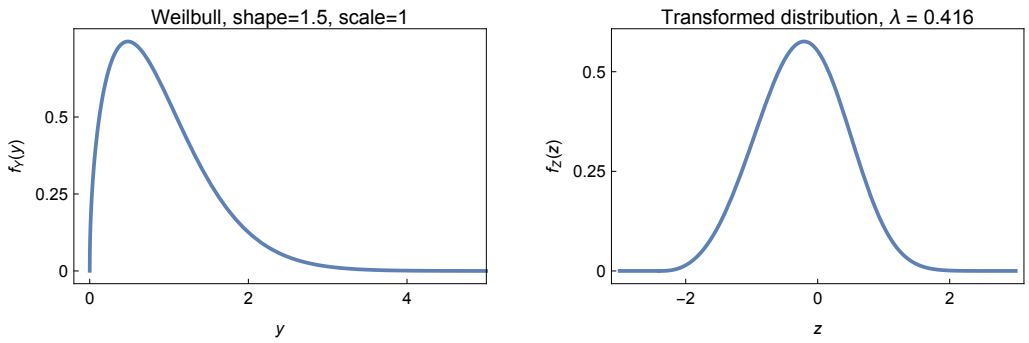

Figure 2.1: Weibull Distribution and a Box-Cox normal approximation.

The inverse or back-transform, $Z \longrightarrow Y$ of the Box-Cox transformation may be written,

$$Y = \begin{cases} (\lambda Z + 1)^{1/\lambda}, & \lambda \neq 0, \\ \exp(Z), & \lambda = 0. \end{cases} \tag{2.3}$$

Box-Cox analysis [Box and Cox, 1964] proceeds by assuming the conditional distribution of the transformed variable $Y^{(\lambda)}$ is normally distributed. eqn. (2.3) requires $\lambda Y^{(\lambda)} + 1 > 0$ when $\lambda \neq 0$. Hence we may conclude that $Y^{(\lambda)}$ has a truncated normal distribution. In the literature this truncation has been entirely ignored or assumed that it's effect is negligible.

In this thesis we will demonstrate that this truncation may have an important effect.

2.1.1 The Box-Cox Distributions

In order to develop an exact treatment of Box-Cox analysis, we start by assuming the following data generation model (DGM). The DGM assumes there is a latent distribution in a sense similar to statistical models for censoring or missing values. In this case the latent distribution is a normal distribution that we refer to as the *Box-Cox Normal Distribution*. This distribution generates Z and then the observed data is generated by the inverse Box-Cox transformation, eqn. (2.3), $Z \rightarrow Y$. The distribution of Y is referred to as the *Box-Cox Data Distribution*. It is a non-normal positive valued distribution which is implicitly defined by the distribution of Z.

2.1.2 Box-Cox Normal Distribution

The PDF for Z is proportional to the normal density $\phi(Z, \mu, \sigma)$ and the constant of proportionality is determined so the density integrates to 1 over the $\left(-\lambda^{-1}, \infty\right)$. Hence the density function for Z when $\lambda > 0$,

$$\phi_{(-\lambda^{-1},\infty)}(z, \mu, \sigma) = \frac{\phi(z, \mu, \sigma)}{1 - \Phi(\xi)}. \tag{2.4}$$

where $\xi = -\left(\lambda^{-1} + \mu\right)\Big/ \sigma$.

Similarly, when $\lambda < 0$,

$$\phi_{(-\infty,-\lambda^{-1})}(z, \mu, \sigma) = \frac{\phi(z, \mu, \sigma)}{\Phi(\xi)}. \tag{2.5}$$

For the general case the Box-Cox Normal Distribution is, $\phi(z, \mu, \sigma, \lambda)$, by

$$\phi(z, \mu, \sigma, \lambda) = \begin{cases} \phi_{(-\lambda^{-1}, \infty)}(z, \mu, \sigma), & \lambda > 0, \\ \phi(z, \mu, \sigma), & \lambda = 0, \\ \phi_{(-\infty, -\lambda^{-1})}(z, \mu, \sigma), & \lambda < 0. \end{cases} \tag{2.6}$$

Standard Box-Cox analysis [Box and Cox, 1964] assumes the transformed data $Z = Y^{(\lambda)}$ is normally distributed $\phi(z, \mu, \sigma)$ and we will refer to this distribution as the Box-Cox normal approximation. The right panel in Figure 2.2 compares the exact and Box-Cox normal approximation when $\lambda = 1$, $\mu = 0$ and $\sigma = 1$.

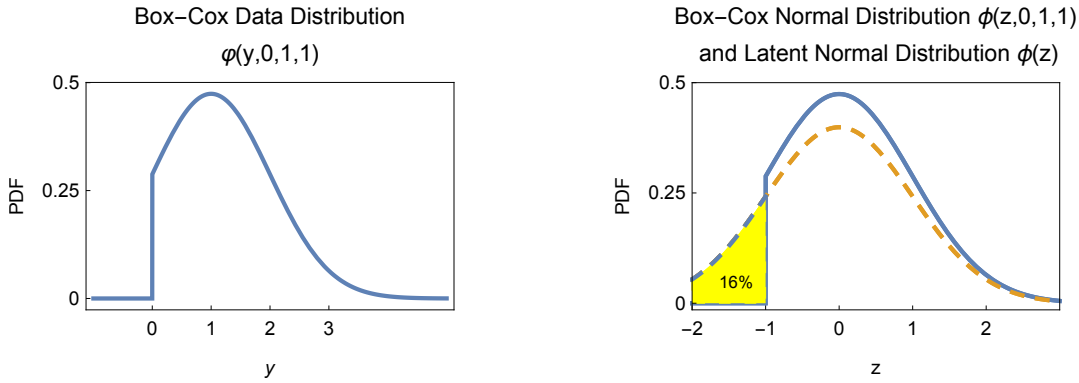

Figure 2.2: Box-Cox Distributions with parameters $\lambda = 1$, $\mu = 0$ and $\sigma = 1$. The exact Box-Cox normal distribution is a truncated normal distribution and its normal approximation distribution. The corresponding Box-Cox data distribution defined by the inverse Box-Cox transformation always has support on $(0, \infty)$.

2.1.3 Kullback-Leibler Divergence

Box-Cox analysis [Box and Cox, 1964] ignores the effect of truncation and so assumes the distribution is normal, $\phi(z, \mu, \sigma)$, $-\infty < z < \infty$. The Kullback-Leibler (KL) divergence may be used to quantify the difference in terms of entropy between the approximate distribution and the exact true truncated distribution.

The KL divergence of the normal approximate distribution from the exact truncated normal distribution when $\lambda > 0$ is given by

$$\begin{aligned} K &= \int_{-\lambda^{-1}}^{\infty} \left(\log \frac{\phi(z, \mu, \sigma)/(1 - \Phi(\xi))}{\phi(z, \mu, \sigma)} \right) \frac{\phi(z, \mu, \sigma)}{1 - \Phi(\xi)} dz \\ &= -\frac{\log(1 - \Phi(\xi))}{1 - \Phi(\xi)} \int_{-\lambda^{-1}}^{\infty} \phi(z, \mu, \sigma) dz \\ &= -\log(1 - \Phi(\xi)), \end{aligned} \tag{2.7}$$

where

$$\xi = -\frac{\lambda^{-1} + \mu}{\sigma}. \tag{2.8}$$

And similarly when $\lambda < 0$, $K = -\log \Phi(\xi)$. When $\lambda > 0$, large $-\xi$ corresponds to $K \doteq 0$ whereas for $\lambda < 0$, large values of ξ correspond $K \doteq 0$.

Let κ be the probability that the back-transform using the Box-Cox normal approximation is valid, that is, $\lambda Z + 1 > 0$. Then we have,

$$\kappa = \begin{cases} 1 - \Phi(\xi), & \lambda > 0, \\ 1, & \lambda = 0, \\ \Phi(\xi), & \lambda < 0. \end{cases} \tag{2.9}$$

Hence for the KL divergence K, we have $\kappa = -e^K$.

In Figure 2.2, $1 - \kappa = \Phi(-1) \doteq 16\%$. This means that if the Box-Cox normal approximation was used to simulate data, it would fail about 16% of the time. In a simple situation involving only independent identical distributions, simply rejecting the invalid data would be an expedient solution but if $1 - \kappa$ was close larger even this approach might not be feasible in more complicated models such as in regression and time series.

The exact Box-Cox Normal Distribution may also be written

$$\phi(z, \mu, \sigma, \lambda) = \kappa^{-1} \phi(z, \mu, \sigma), \quad z \in \mathcal{R}_\lambda, \tag{2.10}$$

where $\mathcal{R}_\lambda$ defines the feasible region for λ in the transformed domain,

$$\mathcal{R}_\lambda = \begin{cases} z \in (-\lambda^{-1}, \infty), & \lambda > 0, \\ z \in (-\infty, \infty), & \lambda = 0, \\ z \in (-\infty, -\lambda^{-1}), & \lambda < 0. \end{cases} \tag{2.11}$$

From Figure 2.3 we conclude that the accuracy improves as $\xi \longrightarrow -\infty$ when $\lambda > 0$ and when $\xi \longrightarrow \infty$ when $\lambda < 0$. It is usually assumed that $\lambda = 1$ indicates no transformation is needed, but strictly speaking we need to also assume that $\xi \ll -3$ so that the truncation effect can be neglected.

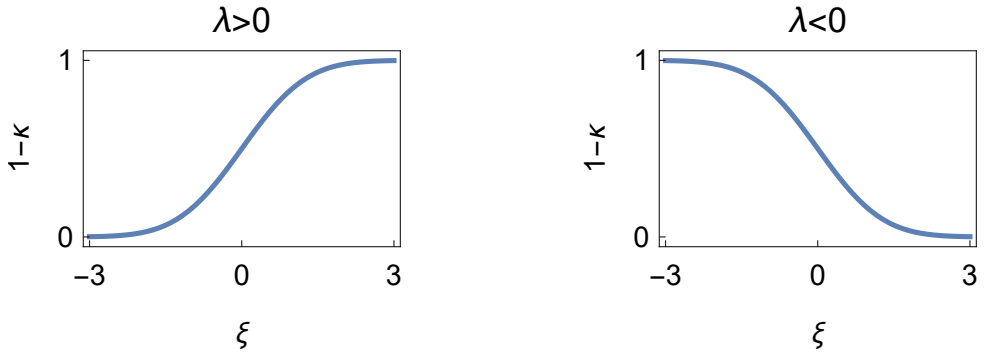

Figure 2.3: Plot of $1 - \kappa$, the probability that the inverse Box-Cox transformation is invalid when the Box-Cox approximation is used, vs. ξ, the standardized truncation limit.

2.1.4 Box-Cox Data Distribution

The distribution that generates the observations $y_1, \ldots, y_n$ is needed to construct the likelihood function. The Box-Cox Normal Distribution, $\phi(z, \mu, \sigma, \lambda)$, defines the distribution of the data in the transformed domain. Applying the inverse Box-Cox transformation $Z \to Y$, the distribution of the data in the original domain is defined. This distribution is denoted by $\varphi(y, \mu, \sigma, \lambda)$ and is defined as the Box-Cox Data Distribution. Using eqn. (2.3) for the inverse transformation to substitute for z and using the Jacobian of the transformation, $d_y z = y^{\lambda-1}$, the *exact* Box-Cox Data Distribution may be written,

$$\varphi(y, \mu, \sigma, \lambda) = \begin{cases} \phi_{(-\lambda^{-1},\infty)}\left(\left(y^\lambda - 1\right)\Big/ \lambda, \mu, \sigma\right) y^{\lambda-1}, & \lambda > 0, \\ \phi(\log(y), \mu, \sigma) y^{-1}, & \lambda = 0, \\ \phi_{(-\infty,-\lambda^{-1})}\left(\left(y^\lambda - 1\right)\Big/ \lambda, \mu, \sigma\right) y^{\lambda-1}, & \lambda < 0. \end{cases} \tag{2.12}$$

Since the Box-Cox distribution is only defined for $Y > 0$, the distribution $\varphi(y, \mu, \sigma, \lambda)$ has support on $(0, \infty)$ as required. To check this note that when $\lambda > 0$, $(y^\lambda - 1)/\lambda \in (-\lambda^{-1}, \infty)$ if and only if $y > 0$. Similarly for $\lambda < 0$. The special case, $\varphi(y, \mu, \sigma, 0)$, is the log-normal distribution.

Box and Cox [1964] implicitly assume for the likelihood computation the approximation that $\varphi(y, \mu, \sigma, \lambda) \doteq \phi(y^{(\lambda)}) y^{\lambda-1}$. For data arising in applications, $y > 0$, so both the Box-Cox transformation and its inverse are valid. In computational statistical inference, such as the parametric bootstrap, this approximation may fail because the inverse Box-Cox transformation may be invalid when the Box-Cox approximate normal distribution is used.

2.2 Simulation of the Box-Cox Data Distribution

Simulation of data from the *Box-Cox Data Distribution* has important applications. In computational statistical inference, methods such as non-parametric bootstrapping, Monte-Carlo testing, Monte-Carlo Markov Chain and cross-validation where it is necessary to simulate data from a fitted hypothetical model. In civil engineering, synthetic or simulated data from empirical models are used for the engineering design. Since Box-Cox models are used in all these applications, it is essential to be able to reliably simulate data from fitted models.

The naive simulation method is simply to generate data from the complete normal distribution and then use the inverse Box-Cox transformation to simulate the data in the untransformed domain. But as we have shown, there is a non-zero probability, $1 - \kappa$ that inverse transformation may be undefined.

In general, the inverse CDF method may be used to generate data from the *Box-Cox Normal Distribution*. Then the Box-Cox transformation is employed to generate the random variates from the *Box-Cox Data Distribution*. This two-step method is usually necessary because the inverse CDF method can not usually be applied directly to the *Box-Cox Data Distribution*.

When $\lambda > 0$, the distribution of Z, $\phi_{(-\lambda^{-1},\infty)}(z, \mu, \sigma)$ is a left-truncated normal with parameters μ, σ and truncation point, $-\lambda^{-1}$. Our objective is to provide an algorithm to simulate Z using an inverse CDF method that transforms a uniform(0,1) random variable to Z making use of efficient algorithms to compute the quantile function from the normal distribution. The quantile function may be expressed as the inverse of the normal CDF and denoted by $\Phi^{-1}(U, \mu, \sigma)$.

The CDF for Z when $\lambda > 0$ may be written,

$$\begin{aligned}\Pr\{Z < z\} &= \int_{-\infty}^{y} \phi_{(-\lambda^{-1},\infty)}(y,\mu,\sigma)\,dy \\ &= \frac{1}{1-\Phi(\xi)} \int_{-\lambda^{-1}}^{z} \phi(z)\,dz \\ &= \frac{\Phi(z,\mu,\sigma) - \Phi\left(-\lambda^{-1},\mu,\sigma\right)}{1-\Phi(\xi)} \\ &= \frac{\Phi(z,\mu,\sigma) - \Phi(\xi)}{1-\Phi(\xi)}.\end{aligned} \tag{2.13}$$

Let U be uniform(0,1) and setting,

$$U = \frac{\Phi(Z,\mu,\sigma) - \Phi(\xi)}{1-\Phi(\xi)}, \tag{2.14}$$

and then simplifying to obtain

$$\Phi(\xi) + U(1-\Phi(\xi)) = \Phi(Z,\mu,\sigma). \tag{2.15}$$

Hence Z may be generated from

$$Z = \Phi^{-1}(U + \Phi(\xi)(1-U),\mu,\sigma). \tag{2.16}$$

Similarly when $\lambda < 0$ we use $Z = \Phi^{-1}(\Phi(\xi)(1-U))$.

When $\lambda = 0$, Z is generated from the normal distribution, $Z = \Phi^{-1}(U,\mu,\sigma)$.
To summarize, to simulate from the *Box-Cox Data Distribution*, we first simulate from the *Box-Cox Normal Distribution* using the equation,

$$Z = \begin{cases} \Phi^{-1}(U + \Phi(\xi)(1-U),\mu,\sigma), & \lambda > 0, \\ \Phi^{-1}(U,\mu,\sigma), & \lambda = 0, \\ \Phi^{-1}(\Phi(\xi)(1-U)), & \lambda < 0, \end{cases} \tag{2.17}$$

where $\xi = -\left(\lambda^{-1} + \mu\right)\Big/ \sigma$.

Then we back-transform to generate data from the *Box-Cox Data Distribution*,

$$Y = \begin{cases} (\lambda Z + 1)^{1/\lambda}, & \lambda \neq 0, \\ \exp(Z), & \lambda = 0. \end{cases} \tag{2.18}$$

The code snippet below shows how the simulation of the *Box-Cox Normal Distribution* is implemented in R.

```
sbxcx <- function(n, mu, sig, lambda) {
#n: number simulated variates, mu, sig, lambda: parameters
 stopifnot(length(mu)==1, length(sig)==1,length(lambda)==1,
           length(n)==1, n>0)
```

```
 if (abs(lambda)<1e-6) {
   Y <- rlnorm(n, mu, sig)
 } else {
  U <- runif(n)
  xi <- -(lambda^(-1)+mu)/sig
  if(lambda > 0) {
   Y <- qnorm(U+pnorm(xi)*(1-U), mu, sig)
  } else {
   Y <- qnorm(pnorm(xi)*U, mu, sig)
  }
 Y
 }
}
```

2.2.1 Illustrative Example

As an illustrative example we consider the Box-Cox Normal Distribution, $\phi(z, 0, 1, 3/4)$. This distribution is the same as a truncated normal distribution with support $(-4/3, \infty)$ and it arises when the Box-Cox transformation $\lambda = 3/4$ is applied to data generated by the Box-Cox Data Distribution, $\varphi(y, 0, 1, 3/4)$. Both of these distributions are shown in the Figure 2.4. The frequency with which the back-transform is invalid is $1 - \kappa = \Phi(-4/3) \doteq 9\%$ using eqn. (2.33) and is represented by the yellow region.

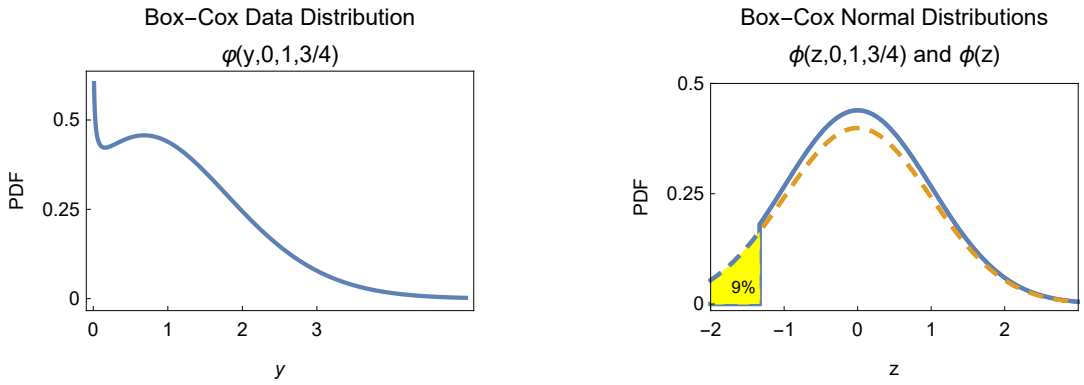

Figure 2.4: Box-Cox Data and Normal Distributions. In the right panel, the dashed curve shows the full normal distribution that is assumed in Box-Cox analysis. The yellow region corresponds to where the back-transform is invalid.

In Figure 2.5, the left panel shows Box-Cox Data Distribution $\varphi(y, 0, 1, 3/4)$ and the probability histogram for 100 random variates generated from this distribution. This histogram resembles data positive data similar to many common distributions and so the right panel shows the corresponding PDF $\phi(z, 0, 1, 3/4)$ and the histogram of the data from the left panel after the Box-Cox transformation with $\lambda = 3/4$.

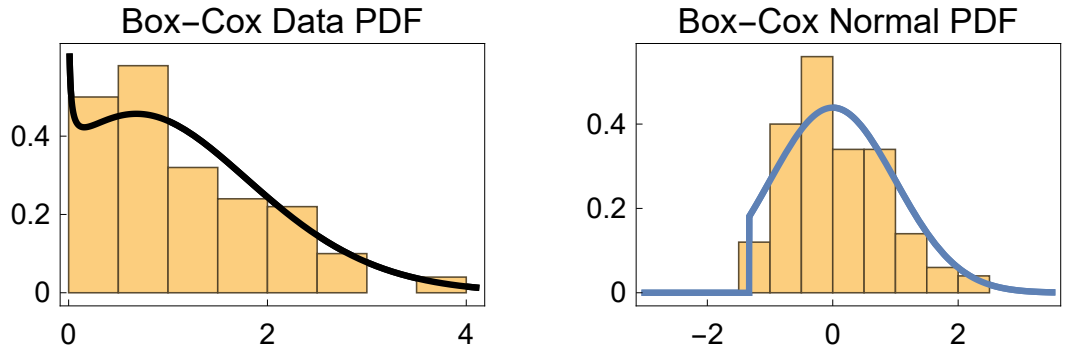

Figure 2.5: Data generated by the distribution $\varphi(y, 0, 1, 3/4)$ is shown in the left panel and the histogram of the transformed data in the right panel.

The data generated in Figure 2.5 was fitted with four widely used distributions. These distributions and their estimated parameters are listed below:

- Exponential Distribution.
$$f(y) = \lambda e^{\lambda(-y)} \tag{2.19}$$
where $\hat{\lambda} \doteq 0.868227$.

- Gamma Distribution.
$$f(y) = \frac{\beta^{-\alpha} y^{\alpha-1} e^{-\frac{y}{\beta}}}{\Gamma(\alpha)} \tag{2.20}$$
where $\hat{\alpha} \doteq 1.32597$ and $\hat{\beta} \doteq 0.868627$

- Weibull Distribution.
$$f(y) = \frac{\alpha e^{-\left(\frac{y}{\beta}\right)^{\alpha}} \left(\frac{y}{\beta}\right)^{\alpha}}{y} \tag{2.21}$$
where $\hat{\alpha} \doteq 1.28821$ and $\hat{\beta} \doteq 1.23387$

- Half-normal Distribution.
$$f(y) = \frac{2\theta e^{-\frac{\theta^2 y^2}{\pi}}}{\pi} \tag{2.22}$$
where $\hat{\theta} \doteq 0.868227$.

In Figure 2.6 the histograms and black curve show the identical data and Box-Cox Data Distribution as in the Figure 2.5 while the blue curve shows the fitted density to this data from the specified distribution. From Figure 2.6 we can conclude that the Box-Cox Data Distribution may generate data similar to that generated by some of the well-known distributions for positive random variables. To investigate this more quantitatively, the Anderson-Darling test was used to see if there is a statistical significant difference. The test statistic, A^2, is defined by eqn. (2.23),
$$A^2 = -\sum_{k=1}^{n} \frac{(2k-1)\left(\log\left(1 - F\left(y_{(-k+n+1)}\right)\right) + \log\left(F\left(y_k\right)\right)\right)}{n} - n, \tag{2.23}$$

where $n = 100$ and $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$ are the sorted data generated from the Box-Cox data distribution $\varphi(y, 0, 1, 3/4)$ that are depicted in the histogram in the left panel of Figure 2.5. The p-value is the for testing hypothesis that the data were generated from each of the specified distributions. Only in the case of the Exponential Distribution is the p-value less than 5%.

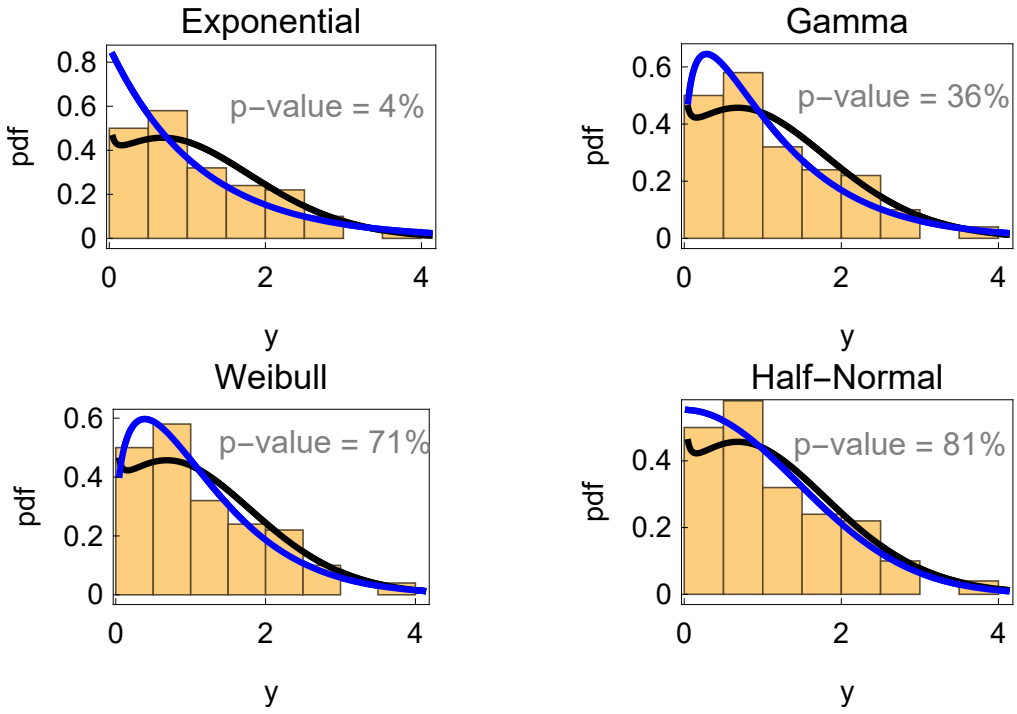

Figure 2.6: Fitting some common Distributions to Box-Cox Data in the left panel of Figure 2.5.

2.3 Exact and Approximate Box-Cox Likelihood Analysis

Often Box-Cox analysis is used with linear models. Suppose there are p explanatory variables and n observations so the Gaussian linear model may be written

$$y_i = \mu_i + \epsilon_i, \tag{2.24}$$

where $\epsilon_i \sim \text{NID}(0, \sigma^2)$ and

$$\mu_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}, \tag{2.25}$$

where for $i = 1, \ldots, n$. We have used the abbreviation $\text{NID}(0, \sigma)$ for normal and independently distributed random variables with mean 0 and standard deviation σ. Box-Cox analysis [Box and Cox, 1964] generalizes eqn. (2.24) by assuming that eqn. (2.24) holds when y_i is replaced by its Box-Cox transform $y_i^{(\lambda)}$ for some λ. Making the transformation $y_i^{(\lambda)} \to y_i$ we can write,

$$y_i = \begin{cases} \left(\lambda y_i^{(\lambda)} + 1\right)^{1/\lambda}, & \lambda \neq 0, \\ \exp(y_i^{(\lambda)}), & \lambda = 0, \end{cases} \tag{2.26}$$

which has Jacobian $y_i^{\lambda-1}$. Hence the PDF for y_i,

$$\tilde{\varphi}(y_i, \mu_i, \sigma, \lambda) = \phi(y_i^{(\lambda)}, \mu_i, \sigma) y_i^{\lambda-1}, \tag{2.27}$$

where we have used the notation $\tilde{\varphi}$ to indicate that this formulation is an approximation that ignores the truncation error since if $y_i^{(\lambda)}$ is normally distributed $\lambda y_i^{(\lambda)} + 1$ has a non-zero probability of being negative which would result in an invalid transformation. As previously noted an invalid transformation does not arise directly in applications since the observed data is always positive.

The log-likelihood using the Box-Cox normal approximation was derived by Box and Cox [1964] and it may be written,

$$\log \tilde{L}(\beta, \sigma, \lambda) = \sum_i \log \phi(y_i^{(\lambda)}, \mu_i, \sigma) + (\lambda - 1) \sum_i \log(y_i), \tag{2.28}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, and $\mu_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}$, $i = 1, \ldots, n$. The likelihood function defined in eqn. (2.28) may be considered as a partial likelihood in the sense of Cox [1975]. It provides an approximation to the complete likelihood defined in eqn. (2.31). The concept of partial likelihood was introduced by Cox [1975] to deal with cases where the complete likelihood was intractable. We will refer to $\tilde{L}(\beta, \sigma, \lambda)$ in eqn. (2.28) as the Partial Box-Cox likelihood. Next we turn to the complete or Exact Box-Cox likelihood which specified in below in eqn. (2.32).

The exact distribution for y_i was given in eqn. (2.12) and is denoted by $\varphi(y_i, \mu_i, \sigma, \lambda)$. An important feature of the method of Box and Cox [1964] is that the computation of maximum likelihood estimate is straightforward because for fixed λ, the standard algorithm for the linear model may be used to obtain the profile likelihood. For fixed λ, the MLE for the other parameters may be obtained by inputting the transformed data to least squares algorithm and hence we obtain, $\tilde{\beta}$, $\tilde{\sigma}$ and $\tilde{\mu}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i,1} + \ldots + \tilde{\beta}_p x_{i,p}$ for $i = 1, \ldots, n$. The Profile Partial Box-Cox log-likelihood

$$\log \tilde{L}_p(\lambda) = \sum_i \log \phi(y_i^{(\lambda)}, \tilde{\mu}_i, \tilde{\sigma}) + (\lambda - 1) \sum_i \log(y_i), \tag{2.29}$$

may be evaluated numerically and plotted over a suitable interval.

Exact Box-Cox analysis may proceed in a similar way but due to the fact that in the transformed domain the exact Box-Cox normal distribution is a truncated normal distribution with support depending on λ, an efficient computational algorithm is more complicated.

From eqn. (2.12), the exact Box-Cox log-likelihood may be written,

$$\log L(\beta, \sigma, \lambda) = \sum_i \log \varphi(y_i, \mu_i, \sigma, \lambda). \tag{2.30}$$

When $\lambda = 0$, eqn. (2.31) is the same as eqn. (2.28) but when $\lambda \neq 0$ normal distribution is replaced by the truncated normal distribution corresponding to support on $(-\lambda^{-1}, \infty)$ or $(-\infty, -\lambda^{-1})$ according as $\lambda > 0$ or $\lambda < 0$ respectively. Thus we obtain,

$$\log L(\beta, \sigma, \lambda) = \begin{cases} \sum_i \log \phi_{(-\lambda^{-1}, \infty)}(y_i^\lambda - 1/\lambda, \mu_i, \sigma) + (\lambda - 1) \sum_i \log y_i, & \lambda > 0, \\ \phi(\log y_i, \mu_i, \sigma) + (\lambda - 1) \sum_i \log y_i, & \lambda = 0, \\ \sum_i \log \phi_{(-\infty, -\lambda^{-1})}(y_i^\lambda - 1/\lambda, \mu_i, \sigma) + (\lambda - 1) \sum_i \log y_i, & \lambda < 0. \end{cases} \tag{2.31}$$

Using eqn. (2.33) we can rewrite this as

$$\log L(\beta, \sigma, \lambda) = \sum_{i=1}^{n} \log(\phi(\frac{y_i^{(\lambda)} - \mu}{\sigma})) + (\lambda - 1) \sum_{i=1} \log y_i - n \log \kappa, \tag{2.32}$$

where

$$\kappa = \begin{cases} 1 - \Phi(-(\lambda^{-1} + \mu)/\sigma), & \lambda > 0, \\ 1, & \lambda = 0, \\ \Phi(-(\lambda^{-1} + \mu)/\sigma), & \lambda < 0. \end{cases} \tag{2.33}$$

Chen and Lockhart [1997] and Chen et al. [2002] assumed that the effect of the Box-Cox approximation is asymptotically negligible but from eqn. (2.32) we see that for $\lambda \neq 0$, the term $-n \log \kappa \longrightarrow \infty$ as $n \longrightarrow \infty$ so it is not negligible. Hence the asymptotic results [Chen and Lockhart, 1997, Chen et al., 2002] are not applicable to the regular case where the model parameters μ, σ and λ remain constant.

The MLE are defined by maximizing $\log L(\beta, \sigma, \lambda)$ over the unknown parameters. This optimization may be solved by first solving the optimization problem for a fixed value of λ and then plotting the profile log-likelihood function of λ as in the approximate analysis [Box and Cox, 1964]. But the problem is made more difficult by the fact that the support or range for the truncated normal distribution depends on the unknown parameter λ when $\lambda \neq 0$. Hence, the sufficient conditions for the asymptotic optimality of the MLE are not satisfied [Shao, 1998]. We conclude that even the standard practice of indicating a 95% confidence interval for the Box-Cox estimate of λ is not justified by asymptotic theory.

We will provide two solutions to the optimization problem. First we present the solution to the case where $\mu_i = \mu$ is constant.

2.3.1 Exact Box-Cox Analysis: Constant Mean Case

We first investigate an efficient algorithm for estimating the parameters in the model specified by the exact Box-Cox Data Distribution, eqn. (2.12). Given a random sample, $y_1, \ldots, y_n$ from this distribution the exact Box-Cox likelihood may be written,

$$\begin{aligned} \log L(\mu, \sigma, \lambda) &= \sum_i \log \varphi(y_i, \mu, \sigma, \lambda) \\ &= \begin{cases} \sum_i \log \phi_{(-\lambda^{-1}, \infty)}(y_i^\lambda - 1/\lambda, \mu, \sigma) + (\lambda - 1) \sum_i \log y_i, & \lambda > 0, \\ \phi(\log y_i, \mu, \sigma) + (\lambda - 1) \sum_i \log y_i, & \lambda = 0, \\ \sum_i \log \phi_{(-\infty, -\lambda^{-1})}(y_i^\lambda - 1/\lambda, \mu, \sigma) + (\lambda - 1) \sum_i \log y_i, & \lambda < 0. \end{cases} \end{aligned} \tag{2.34}$$

The MLE are defined by,

$$(\hat{\mu}, \hat{\sigma}, \hat{\lambda}) = \underset{\mu, \sigma, \lambda}{\operatorname{argmax}} \log L(\mu, \sigma, \lambda). \tag{2.35}$$

Most general purpose non-linear optimization routines can not handle this problem since the range of the objective function depends on one of the parameters, viz. λ and as we have already noted the sufficient conditions [Shao, 1998] for the statistical optimality of the MLE are also not satisfied for this very reason.

We can proceed by defining the profile log-likelihood function for λ

$$\log L_p(\lambda) = \underset{\mu,\sigma}{\operatorname{argmax}} \log L(\mu, \sigma, \lambda). \tag{2.36}$$

The profile log-likelihood for λ may then be plotted to obtain a range of suitable values for λ or optimized using an one-dimensional optimization algorithm.

So for a fixed λ our problem reduces to determining the MLE in a truncated normal distribution. Cohen [1991] provides an efficient algorithm for this problem. First an iterative algorithm is developed for obtaining the method of moments estimator for μ and σ in a singly left-truncated normal distribution (T, ∞). These estimates can then be used as starting values in another iterative algorithm to obtain the MLE's. The algorithm for right-truncated case may be obtained by simply negating the data and treating it as left-truncated.

Cohen Algorithm for the Truncated Normal Distribution

We will assume that Z has a left-truncated normal distribution with truncation point $T = -\lambda^{-1}$ and the normal parameters are μ and σ. Let W be the corresponding standardized variable $W = (Z-\mu)/\sigma$ and let $\xi = (T-\mu)/\sigma$ be the corresponding standardized truncation point. Then the kth moment about zero of W can be written [Cohen, 1991],

$$\mu_w(k) = \frac{1}{1 - \Phi(\xi)} \int_{\xi}^{\infty} t^k \phi(t) dt, \tag{2.37}$$

where $\Phi(\xi)$ is the standard normal CDF evaluated at ξ and $\phi(t)$ is the standard normal density function. The first two moments may be obtained using integration-by-parts. For the mean of W, $k = 1$ in eqn. (2.37) and we have,

$$\mu_w(1) = Q \iff E\{(Z - \mu)/\sigma\} = Q, \tag{2.38}$$

where

$$Q = \frac{\phi(\xi)}{1 - \Phi(\xi)}. \tag{2.39}$$

Hence the mean of Z are given by,

$$E\{Z\} = \mu + \sigma Q. \tag{2.40}$$

Similarly for the variance of W,

$$\mu_w(2) = 1 + \xi Q. \tag{2.41}$$

$$\begin{aligned} \operatorname{Var}(W) &= \tilde{\mu}_2 - \tilde{\mu}_1^2 \\ &= 1 + \xi Q - Q^2 \\ &= 1 + Q(\xi - Q) \\ &= \operatorname{Var}((Z - \mu)/\sigma) \\ &= 1/\sigma^2 \operatorname{Var}(Z). \end{aligned} \tag{2.42}$$

Hence,

$$\text{Var}\{Z\} = \sigma^2(1 - Q(Q - \xi)). \tag{2.43}$$

As a numerical check when $T = -1/1.75$, $\mu = 0$ and $\sigma = 1$, we find $Q = 0.473155 = E\{Z\}$ and $\text{Var}(Z) = 0.50575$. These values agree with Mathematica's methods for truncated distributions.

eqn. (2.40) and (2.43) can be solved for the normal parameters μ and σ given the sample estimates for $E\{Z\}$ and $\text{Var}(Z)$ using an iterative algorithm such as Newton's method [Ortega and Rheinboldt, 2000] which is implemented in the Mathematica function FindRoot[]. The method-of-moments, obtained using this method are $\tilde{\mu} = -0.326428$ and $\tilde{\sigma} = 1.12147$. The code snippet shows how this is implemented in Mathematica. Verifying that this is indeed the solution to the moment equations, we find that it is correct.

Let we assume,

$$s_z^2 = \sigma^2(1 - Q(Q - \xi)), \tag{2.44}$$

and

$$\bar{z} = \mu + \sigma Q. \tag{2.45}$$

Cohen [1991] presented a simpler algorithm replaces the need for solving a non-linear system of two equations by only solving a single non-linear equation which is algorithmically much simpler and greatly easier to implement in programming environments such as R. Let $\bar{z}$ and s^2 be the sample mean and variance given data $z_1, \ldots, z_n$. Then the estimating equations derived from eqn. (2.40) and (2.43) may be written, $s^2 = \sigma^2(1-Q(\xi)(Q(\xi)-\xi))$ and $\bar{z}-T = \sigma(\xi-Q(\xi))$. Hence,

$$\begin{aligned} \frac{s^2}{(\bar{z} - T)^2} &= \frac{\sigma^2[1 - Q(\xi)(Q(\xi) - \xi)]}{\sigma^2[\xi - Q(\xi)]^2} \\ &= \frac{[1 - Q(\xi)(Q(\xi) - \xi)]}{[\xi - Q(\xi)]^2} = \alpha(\xi), \end{aligned} \tag{2.46}$$

where $Q(\xi) = \phi(\xi)/(1 - \Phi(\xi))$. eqn. (2.46) may be solved by an one-dimensional optimization algorithm. Verification of our equation demonstrated as follows,

$$\frac{s^2}{(\bar{z} - T)^2} = \frac{1 - Q(\xi)(Q(\xi) - \xi)}{(Q(\xi) - \xi)^2}. \tag{2.47}$$

Continuing with our numerical example, $\bar{x}$ and s^2 are the sample mean and the sample variance respectively in the data and also T is truncated point given. Therefore, the method of moments estimators using FindRoot[] are {-0.326428,1.12147}.

The maximum likelihood equations are somewhat complex. Taking the first derivatives,

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} (y_i^{(\lambda)} - \mu)/\sigma^2 - n(1/\sigma)\frac{\phi(\xi)}{(1 - \Phi(\xi))}, \tag{2.48}$$

and the first derivatives with respect to σ^2 is given by,

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} (y_i^{(\lambda)} - \mu)^2 - n(\frac{1}{2\sigma^2})\left(\frac{-(\lambda^{-1} + \mu)}{\sigma}\right)\frac{-\phi(\xi)}{(1 - \Phi(\xi))}, \tag{2.49}$$

then it implies from eqn. (2.49),

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i^{(\lambda)} - \mu)^2 + \frac{n}{2\sigma^2}\xi\Psi(\xi). \tag{2.50}$$

By setting $\partial l/\partial\mu = 0$, we have the first MLE as follows,

$$\bar{y}^{(\lambda)} = \mu + \sigma\Psi(\xi), \tag{2.51}$$

where $\Psi(.) = \phi(.)/(1 - \Phi(.))$ and $Z = y_i^{(\lambda)}$ denoted as transformed variables. This equation determines that in the case of $\lambda > 0$, the mean of left truncation, $\bar{y}^{(\lambda)}$ is approximately equivalent to true population mean μ plus a part which depends on the truncation point ξ and σ. Then, let to set $\partial l/\partial\sigma^2 = 0$ and simplify,

$$n\hat{\sigma}^2_{y^{(\lambda)}} + n(\bar{y}^{(\lambda)})^2 = n\sigma^2\xi\Psi(\xi) + n\sigma^2 + n\mu(2\bar{y}^{(\lambda)} - \mu). \tag{2.52}$$

So, we have

$$\hat{\sigma}^2_{y^{(\lambda)}} = \sigma^2\left[1 - \Psi(\xi)(\Psi(\xi) - \xi)\right], \tag{2.53}$$

where $\hat{\sigma}^2_{y^{(\lambda)}} = 1/n\sum_{i=1}^{n}\left(y^{(\lambda)} - \bar{y}^{(\lambda)}\right)^2$. Cohen [1950, 1991] indicated that the maximum likelihood estimations can be computed by these equations iteratively. However, convergency can not always be obtained by this method due to dependency of convergence on the initial values. Cohen [1949, 1950] suggested that the initial values should be the sample moments in the Newton-Raphson method. Pearson and Lee [1908] considered the method of moment estimation to find μ and σ for truncated normal distribution.

Simulated Example

We compare the exact and profile partial log-likelihoods using the simulated example discussed in Section 2.2.1. Recall that in this case, the Box-Cox Data Distribution defined by $\varphi(y, \mu = 0, \sigma = 1, \lambda = 3/4)$. In this example $1 - \kappa = 9\%$ so the truncation effect is sizable but realistic for various positively skewed distributions that may be seen in practice as was demonstrated in the previous section.

Data generated was from samples with $n = 100$ or $n = 1000$ and the profile exact log-likelihoods are shown in the two left panels in Figure 2.7. These two panels show that as n increases the accuracy of the estimate for λ improves. Not only does the MLE estimate itself change from $\hat{\lambda} = 0.67$ when $n = 100$ to $\hat{\lambda} = 0.71$ when $n = 1000$ but the profile log-likelihood becomes more concentrated about the true value $\lambda = 0.75$ and the width of the 95% confidence interval decreases from about (0.4, 1.2) in the top left panel to about (0.6, 0.8) in the bottom left panel.[1]

This is not the case for the Box-Cox profile partial likelihood. There is no improvement in accuracy and the right panels in Figure 2.7 suggest the estimate may not even be consistent.

[1] As we have noted asymptotic theory is not available to justify these confidence intervals. But this interval may also be interpreted as a 38% plausibility interval in the sense of relative likelihood [Sprott, 2000, 2.4]. To see this note that the 95% confidence interval is determined by $-2(\log L_m - \log L_0) = 1.92$ where $\log L_m$ is the maximized log-likelihood and $\log L_0$ determines the confidence interval endpoints. Hence, $L_m/L_0 = 38\%$.

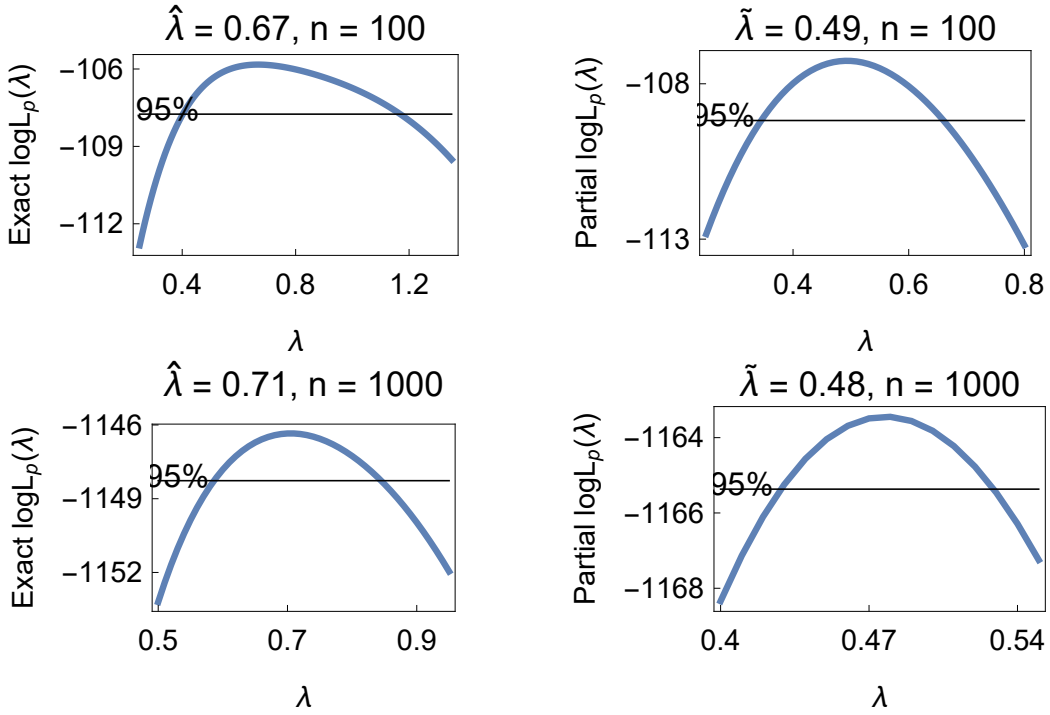

Figure 2.7: Exact and approximate likelihood analysis of simulated data from a Box-Cox data distribution with parameters $\mu = 0$, $\sigma = 1$, $\lambda = 0.75$ for sample sizes $n = 100$ and $n = 1000$.

Application to Length of rivers dataset

The built-in R dataset 'rivers' is comprised of the lengths of the longest 141 rivers in miles in North America. A histogram with for this dataset is shown in Figure 2.8 and we note the data exhibit strong positive skewness with a long right tail.

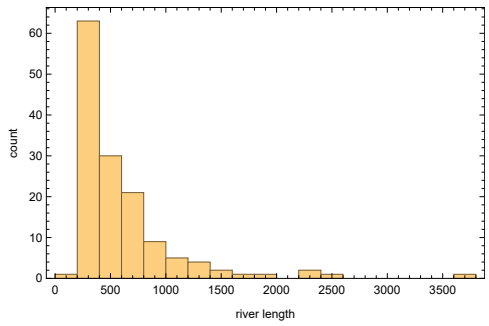

Figure 2.8: Histogram of 'rivers' dataset.

The exact and approximate Box-Cox analyzes are shown in Figure 2.9 and we see that the MLE's for λ are in close agreement although the shape of the likelihood function differs so the statistical inference is not the same. The exact likelihood produces an asymmetric 95% confidence interval $(-1.05, -0.3)$ while for the approximate method the interval is $(-0.81, -0.3)$. It is surprising that the inference changes this much for this data since the mean and standard

deviation of the data in the transformed domain are respectively 1.75432 and 0.0184787 so $\hat{\xi} = 3.45$ and $1 - \hat{\kappa} = 0.0003$.

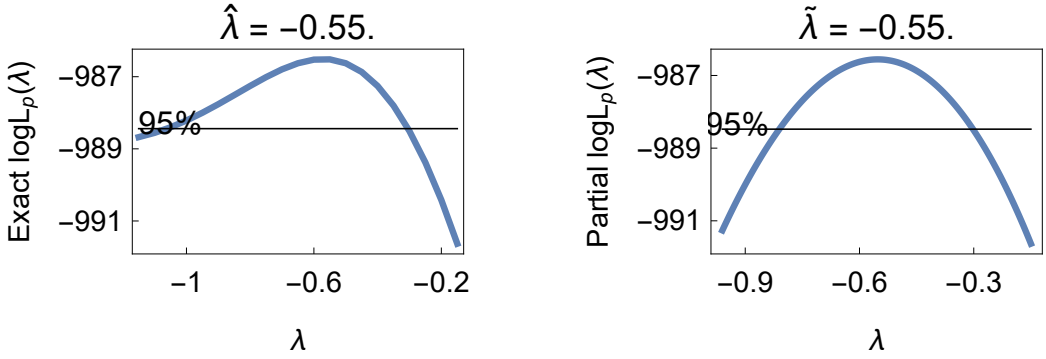

Figure 2.9: Exact and approximate likelihood analysis of 'rivers' dataset.

2.3.2 Exact Box-Cox Analysis: Regression Case

In the more general regression case, we have a response variable Y and p explanatory variables $X_1, \ldots, X_p$ and Y is related to the inputs through the regression equation,

$$\mu_x = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, \tag{2.54}$$

where the PDF $\varphi(y, \mu_x, \sigma, \lambda)$ specifies the statistical model.

Given n observations and the data $(x_{i,1}, \ldots, x_{i,p}, y_i)$, $i = 1, \ldots, n$ the log-likelihood function is determined by the Box-Cox Data Distribution, eqn. (2.12),

$$\log L(\beta, \sigma, \lambda) = \sum_{i=1}^{n} \log \varphi(y_i, \mu_i, \sigma, \lambda), \tag{2.55}$$

where

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}. \tag{2.56}$$

When $\lambda > 0$ the $\log L$ may be directly computed from the simplified result,

$$\log L(\beta, \sigma, \lambda) = \sum_{i=1}^{n} \log \phi(\frac{y_i^{\lambda} - 1}{\lambda}, \mu_i, \sigma, \lambda) + (\lambda - 1) \sum_{i=1}^{n} \log(y_i) - n \log(1 - \Phi(-\frac{1}{\lambda}, \mu_i, \sigma)), \tag{2.57}$$

where μ_i is defined as in eqn. (2.56). As explained in eqn. (2.12), the positivity assumption, $Y_i > 0, i = 1, \ldots, n$ guarantees that eqn. (2.57) is always defined. Furthermore, similar expressions may be given when $\lambda < 0$.
For fixed λ the exact profile log-likelihood is defined by

$$\log L_p(\lambda) = \max_{\beta, \sigma} \log L(\beta, \sigma, \lambda) \tag{2.58}$$

Provided p is not too large, it is feasible to carry out the maximization in eqn. (2.58) using a general purpose optimization.

2.3.3 An Approximation to the Profile Log-likelihood

As an approximation to the profile log-likelihood defined in eqn. (2.58) we consider,

$$\log \bar{L}_p(\lambda) = \log L(\tilde{\beta}, \tilde{\sigma}, \lambda), \tag{2.59}$$

where $\tilde{\beta}$ and $\tilde{\sigma}$ are obtained using the Box-Cox method [Box and Cox, 1964].

We illustrate this method using the same data as used in Figure 2.7 with the constant mean case. The parameter settings are: $\mu = 0$, $\sigma = 1$, $\lambda = 0.75$ with $n = 100$ and $n = 1000$ observations. In Figure 2.7, the exact MLE for λ were $\hat{\lambda} = 0.67$ and $\hat{\lambda} = 0.71$ corresponding to $n = 100$ and $n = 1000$ respectively while the Box-Cox method [Box and Cox, 1964] produced $\tilde{\lambda} = 0.49, 0.48$ corresponding to $n = 100, 1000$. The suggested approximation produces $\bar{\lambda} = 0.46, 0.56$ respectively for $n = 100, 1000$ as shown below in Figure 2.10. In this case, the approximation is not an accurate estimate of the MLE.

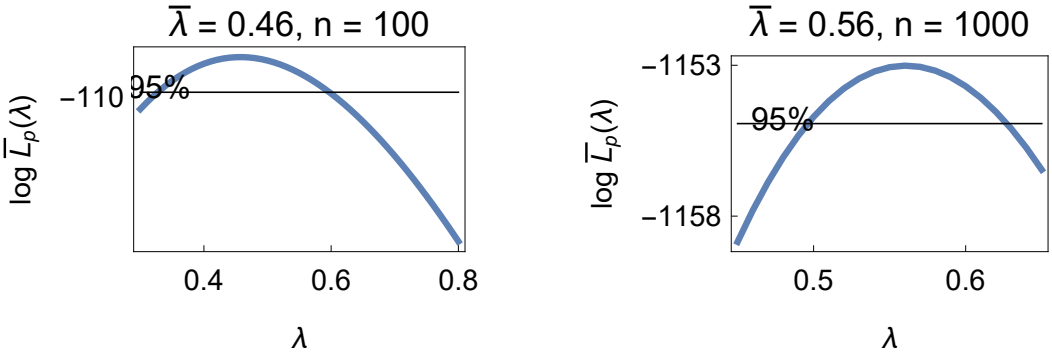

Figure 2.10: Exact and approximate likelihood analysis of simulated data from a Box-Cox data distribution with parameters $\mu = 0$, $\sigma = 1$, $\lambda = 0.75$ for sample sizes $n = 100$ and $n = 1000$.

2.4 EM Algorithm for Truncated Normal Regression

EM algorithm was investigated since it may provide a more efficient algorithm for doing the optimization required for computing the exact profile log-likelihood function given in eqn. (2.58). Initially we start with the simple case of only the parameter $\beta_0 = \mu$. So we consider implementing the EM algorithm for estimating the parameters μ and σ^2 from truncated normal distribution, $\phi(z, \mu, \sigma, \lambda)$. In this section we consider the case $\lambda > 0$, which corresponds to the normal distribution with parameters μ, σ and truncation point $T = -1/\lambda$. Given n observations from this distribution $u_1, \ldots, u_n$, we consider that there are m latent unknown observations $v_1, \ldots, v_m$ corresponding to the missing truncated data where m is determined from $(n+m)(1 - \Phi(T, \mu, \sigma)) = n$ so $m = n\Phi(T, \mu, \sigma)/(1 - \Phi(T, \mu, \sigma))$.

Given initial estimates for the parameters, for example $\mu^{(0)} = n^{-1} \sum_i z_i$ and $\sigma^{(0)}$ the sample standard deviation of $u_1, \ldots, u_n$. Initialize an iteration counter $i \leftarrow 0$.

2.4.1 Expectation step

The mean of the latent sample has expectation μ^*, where μ^* is the expected value from the truncated distribution $\phi_{(-\infty,T)}(\mu^{(i)}, \sigma^{(i)})$ that can be determined from eqn. (2.60) below.

Mean of Truncated Normal

In general, let $\mu^*(\mu, \sigma, T)$ denote the expected value from the truncated distribution with PDF $\phi_{(-\infty,T)}(\mu, \sigma)$ then using Mathematica it can be shown that,

$$E\{Z\} = \sqrt{\frac{2}{\pi}}\sigma e^{-\frac{(\lambda\mu+1)^2}{2\lambda^2\sigma^2}} / \left(1 + \text{erf}\left(\frac{\lambda\mu + 1}{\sqrt{2}\lambda\sigma}\right)\right) \tag{2.60}$$

where erf(t) denotes the error function defined by,

$$\text{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

for $z \geq 0$ and $\text{erfc}(z) = 1 - \text{erf}(z)$.

2.4.2 Maximization step

In this case, the updated estimates are simply determined by

$$\mu^{(i+1)} \leftarrow \frac{n\bar{u} + m\mu^{(*)}}{n + m} \tag{2.61}$$

and

$$\sigma^{(i+1)} \leftarrow \sqrt{(n+m)^{-1}\left[\sum_i (u_i - \mu^{(i+1)})^2 + mE_{\mu^{(i)},\sigma^{(i)}}(Z - \mu)^2\right]} \tag{2.62}$$

2.4.3 Iteration

The steps Expectation and Maximization are iterated until the estimates converge.

2.5 Simulated Example

For parameter settings, $x_i \sim N(0, 2)$, $\beta_0 = 0$, $\beta_1 = 0.5$, $\sigma = 1$, $\lambda = 0.75$ with $n = 100$ and $n = 200$ observations we compare the exact MLE for λ with the Box-Cox estimates in Figures 2.11 and 2.12. The exact MLE were obtained by evaluating the exact profile log-likelihood in eqn. (2.58) over a fine-grid while the Box-Cox estimates were obtained using the R function 'MASS:boxcox()'. The exact MLE is more accurate and its accuracy improves as n increases. As we can seen from Figure 2.11, the true $\lambda = 0.75$ included in the confidence interval. Howevere, estimates of λ obtained by built-in R function are not consistent with initial $\lambda = 0.75$ as shown in Figure 2.12.

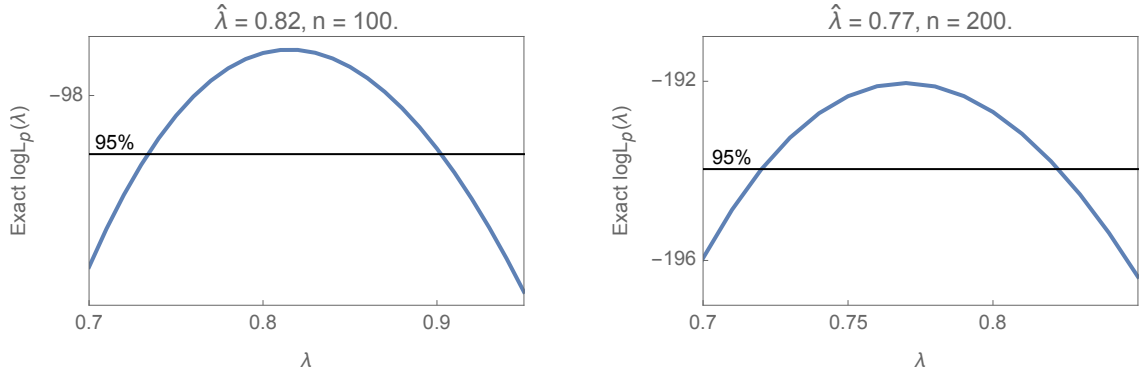

Figure 2.11: Exact Box-Cox analysis with simulated regression with $\lambda = 0.75$ and $\mu_i = \beta_0 + \beta_i x_i$, $i = 1, \ldots, n$

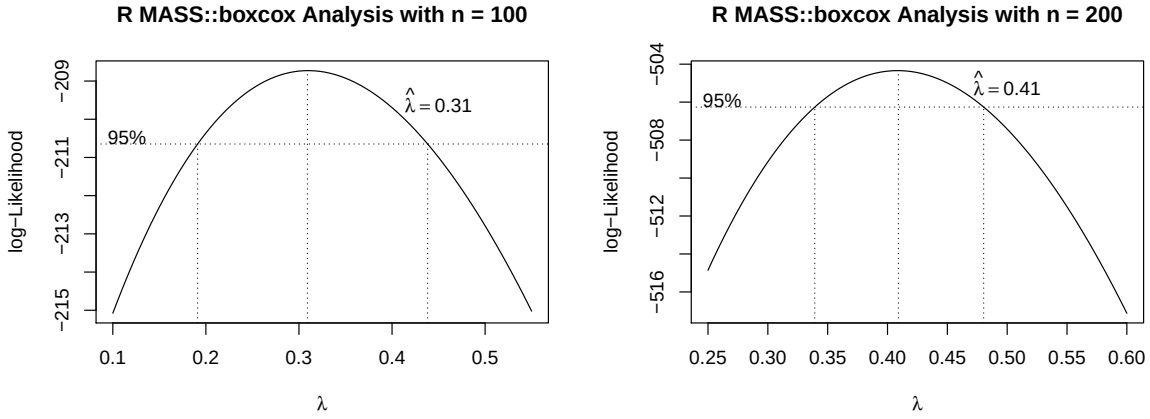

Figure 2.12: Approximate Box-Cox analysis using R with simulated regression with $\lambda = 0.75$ and $\mu_i = \beta_0 + \beta_i x_i$, $i = 1, \ldots, n$

Chapter 3

Box-Cox Time Series

3.1 Introduction

Sampling from a truncated multivariate normal (TMVN) distribution is a recurring problem in many areas in statistics and econometrics, such as Kriging, censored data models, and general linear time series model. In this chapter, we aim to discuss about the simulation of multivariate normal distribution on truncated data.
To simulate univariate truncated normal distribution, classical inverse CDF method can be used as follows,

$$x = \Phi^{-1}\left[\Phi(a) + u\left(\Phi(b) - \Phi(a)\right)\right], \tag{3.1}$$

where $u \sim Unif[0, 1]$ and the boundrary function $[a, b]$. It is denoted Φ as CDF of the normal distribution and Φ^{-1} its inverse.

Marsaglia [1964] introduced an accept-reject algorithm to generate random variables by using the tail of the normal distribution. Later, the Ziggurat algorithm was proposed by Marsaglia and Tsang [2000] for generating random variables used horizontal rectangles. Robert [1995] extended the idea of the previous algorithm proposed by Marsaglia [1964] to simulate univariate truncated normal distribution. Chopin [2011] designed the fast simulation for truncated Gaussian distributions based on the Ziggurat algorithm of Marsaglia and Tsang [2000]. To generate the one-sided truncated variables, it can be naively sampling from a normal distribution untill the desired random variables in the specific interval obtained. Rejection sampling can be considered as a delicate approach since it depends on probabilities of acceptance [Robert, 1995]. For high-dimensional truncated normal distribution, accept and reject approach is not preferable.

Gibbs sampler is regarded as a special case of the Markov chain Monte Carlo (MCMC) algorithm to simulate multivariate truncated normal variables for any covariance structure, specifically when direct sampling is challenging. The Markov chain sequences $x^{(n)}$ can be obtained repeatedly from $TN_p(\mu, \Sigma, x_i^-, x_i^+)$.
The Gibbs sampling procedure is defined by Robert [1995] as follows,

1. $x_1^{(n)} \sim TN\left(E[x_1|x_2^{(n-1)}, ..., x_p^{(n-1)}], x_1^-, x_1^+, \sigma_1^2\right)$,

2. $x_2^{(n)} \sim TN\left(E[x_2|x_1^{(n)}, x_3^{(n-1)}, ..., x_p^{(n-1)}], x_2^-, x_2^+, \sigma_2^2\right)$,

   ...

3. $x_p^{(n)} \sim TN\left(E[x_p|x_1^{(n)}, ..., x_{p-1}^{(n)}], x_p^-, x_p^+, \sigma_p^2\right)$,

where $[x_i^-, x_i^+]$ is truncated interval on p dimensions. The conditional expectation and variances for x_i can be obtained by,

$$E[x_i|x_{-i}] = \mu_i + \Sigma_{i-i}\Sigma_{-i-i}^{-1}(x_{-i} - \mu_{-i}),$$

and

$$\sigma_i^2 = \sigma_{ii}^2 - \Sigma_{i-i}\Sigma_{-i-i}^{-1}\Sigma_{i-i}.$$

Gibbs sampler can be performed for Bayesian calculations of constrained parameter and truncated data problems [Gelfand and Smith, 1990, Gelfand et al., 1992]. Furthermore, Gelfand et al. [1992] illustrated Gibbs sampling application in ordered parameters from exponential distributions and censored data with regression model. Robert [1995] suggests the Gibbs sampler theory for sampling from truncated multivariate normal distribution.

Gibbs sampler was proposed by Chen and Deely [1992] for constrained multiple linear regression. Chen and Schmeiser [1996] showed that Gibbs sampler performs better on independent random variables and reported to sample from conditional distribution by Gibbs algorithm. Rodriguez-Yam et al. [2004] also employed Gibbs sampler algorithm in multiple linear regression with inequality constraints parameters. Sampling from truncated densities introduced by Damien and Walker [2001] in terms of latent variable idea within Gibbs sampling context.

Both rejection and Gibbs sampling algorithm may have disadvantages in the context of computations. Robert [1995] indicated that Gibbs sampling as an efficient and fast algorithm to generate random variables from the truncated multivariate normal distribution. It is important to consider that the convergence of the Gibbs sampling to the stationary distribution may be computationally complex. Moreover, Gelman and Rubin [1992] stated that inferences from iterative simulation would be less efficient compared to direct simulation, even though iterative simulations have wider applications. In fact, sampling from conditional distributions takes more computing time in some cases.

We propose a direct approach to produce the conditional simulation via decomposition of covariance matrix and present the application of Box-Cox transformations in time series model in this chapter. In the case of highly correlated random vector, Cholesky method would be employed to determine the optimal and efficient ones. The alternative methodology can be used for sampling from truncated multivariate normal distribution and providing simulation algorithm to minimize the loss function. The simulation of truncated normal variables need to reconsider in Bayesian inference for some truncated parameter space problems. We further demonstrate that this approach can be applied for any arbitrary covariance structure.

Latter, we aim to generalize these procedures in an applied context and the contribution of exact Box-Cox analysis in simulation. In this chapter, Cholesky decomposition and Durbin-Levinson are both employed to generate Box-Cox transformed time series and its inverse. Furthermore, the forecasting and simulation of time series would be investigated and the results are compared with the Box-Cox normal approximation.

3.2 Truncated Multivariate Normal Distribution

Let Z be a random vector such that $Z = (Z_1, Z_2)$ where Z_1 and Z_2 have a jointly normal distribution as written by,

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right).$$

The conditional distribution of Z_2 given $Z_1 = z_1$ has multivariate normal distribution with mean,

$$\mu_{Z_2|Z_1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Z_1 - \mu_1), \tag{3.2}$$

and the covariance matrix

$$\Sigma_{Z_2|Z_1} = \Sigma_{22} + \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \tag{3.3}$$

Direct method was used for generating Box-Cox transformed stationary time series. As we mentioned in Chapter 2, the Box-Cox family distribution can be simulated by truncated normal distribution.

3.3 Simulation of Truncated Normal Variables

3.3.1 Bivariate case

We initially consider the following example presented by Robert [1995]. Sampling from the truncated normal distribution with truncation space $\mathfrak{R}$ can be written as,

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim TN\left(0, \Sigma, \mathfrak{R}\right),$$

where

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

and A is the circle of center $\gamma = (\gamma_1, \gamma_2)$ and radius r. Hence, it can be defined by,

$$x_1^-(x_2) = \gamma_1 - \sqrt{r^2 - (\gamma_2 - x_2)^2},$$

$$x_1^+(x_2) = \gamma_1 + \sqrt{r^2 - (\gamma_2 - x_2)^2},$$

and

$$x_2^-(x_1) = \gamma_2 - \sqrt{r^2 - (\gamma_1 - x_1)^2},$$

$$x_2^-(x_1) = \gamma_2 + \sqrt{r^2 - (\gamma_1 - x_1)^2}.$$

The truncated bivariate normal distribution is obtained from Gibbs sampling as follows,

1. $x_1^{(n)} \sim TN(\rho x_2^{(n-1)}, x_1^-(x_2^{(n-1)}), x_1^+(x_2^{(n-1)}), 1 - \rho^2)$,

2. $x_2^{(n)} \sim TN(\rho x_1^{(n)}, x_2^-(x_1^{(n)}), x_2^+(x_1^{(n)}), 1-\rho^2)$.

In this study, we propose a general and direct approach for simulation of a bivariate truncated normal distribution. Further, it will be illustrated its contribution in simulating Box-Cox transformations. Let Σ be the desired covariance matrix which produces the random number. Assuming Σ is symmetric and positive-definite and can be decomposed by Cholesky transformation where $\Sigma = LU = LL^T$. Let L be a lower triangle matrix, therefore we have,

$$Y = LZ. \tag{3.4}$$

We assume that Z_1 and Z_2 are uncorrelated random variables from a truncated normal distribution. In first case, we create a realization of bivariate distribution $Y = (Y_1, Y_2)$ by using Cholesky factorization as shown,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}. \tag{3.5}$$

Then, it can be expressed as,

$$\begin{aligned} L_{11}Z_1 &= Y_1, \\ L_{21}Z_1 + L_{22}Z_2 &= Y_2. \end{aligned} \tag{3.6}$$

By considering $Z = (Z_1, Z_2)$ be an independent vector right-truncated at point $b = (b_1, b_2)$. Hence, the main purpose is to investigate the boundary in Y-space when we investigate as written $Z \to Y$,

$$\begin{aligned} Z_1 &\leq b_1, \\ Z_2 &\leq b_2. \end{aligned} \tag{3.7}$$

and we have,

$$\begin{aligned} Y_1 &\leq L_{11}b_1, \\ Y_2 &\leq L_{22}b_2 + L_{21}Z_1. \end{aligned} \tag{3.8}$$

The algorithm that can be used for the simulation of truncated normal variables consists in the following steps:

- **Step 1:** Generate two uncorrelated random vectors Z_1 and Z_2 from the Box-Cox Normal Distribution with non-zero mean μ and variance σ^2.
- **Step 2:** Given the coefficients and autocovariance function.
- **Step 3:** Calculate the lower triangular matrix of covariance matrix, $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.
- **Step 4:** Use the lower triangular matrix $\Sigma = LL^T$.
- **Step 5:** Obtain the one-sided truncated normal distribution.

The application of the this method for bivariate truncated normal distribution can be shown in Figures 3.1 amd 3.2. We plot in Figure 3.1 a random sample of a size $n = 10,000$ generated from a truncated normal distribution with truncation point $-\lambda^{-1}$ and the mean $\mu = 2$ and standard deviation $\sigma = 1$. Figure 3.2 illustrates the simulated random variables from bivariate truncated normal with correlation $\rho = 0.9$ and $\lambda = 0.5$.

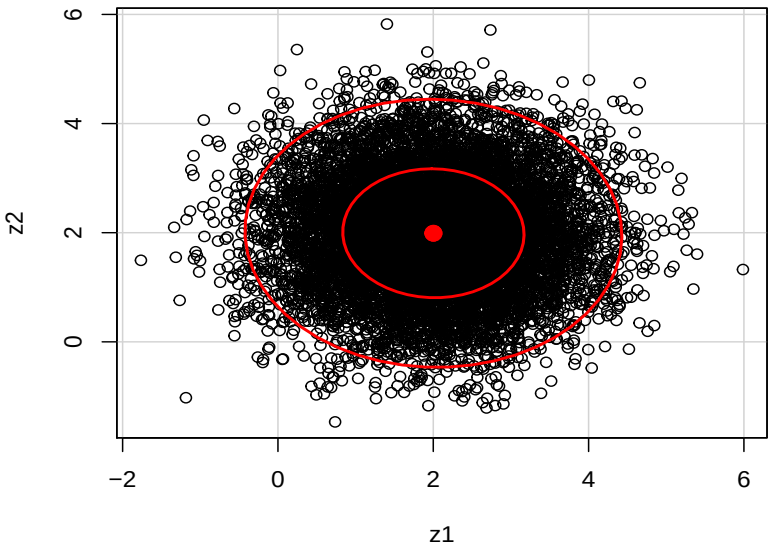

Figure 3.1: Ellipsoids of concentration corresponding to 0.95 and 0.5 probability for simulated random variables from Box-Cox distribution with $\lambda = 0.5$, $\mu = 2$ and $\sigma = 1$.

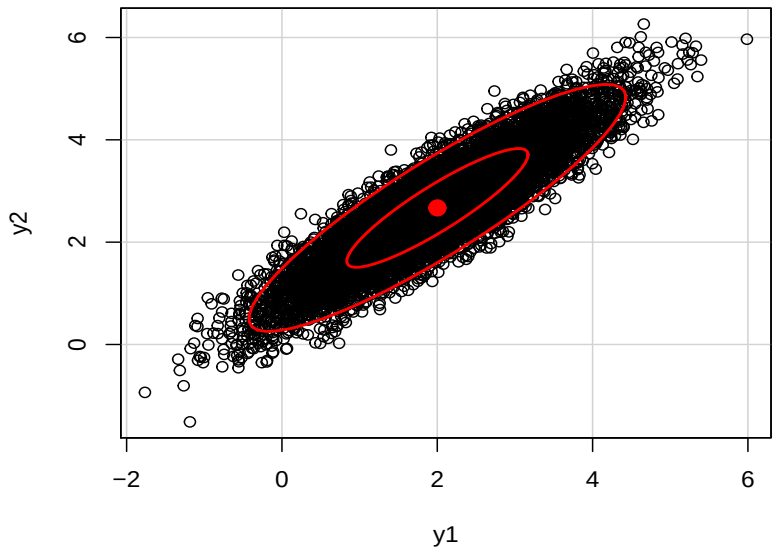

Figure 3.2: Ellipsoids of concentration corresponding to 0.95 and 0.5 probability for simulated random variables from Box-Cox distribution with $\lambda = 0.5$ and $\rho = 0.9$ $\mu = 2$ and $\sigma = 1$.

3.3.2 Multivariate case

We generate $Y_1, \ldots, Y_p$ from $TN_p(\mu, \sigma^2, a_i, b_i)$ and derive the Cholesky decomposition of the covariance matrix, Σ, then use of a lower triangular matrix L to simulate truncated normal distribution with desired covariance matrix. Denote Y_i be an independent vector with right truncation point $b_i = (b_1, ..., b_p)$. To determine the Z-boundaries of a given space, Y_i can be easily computed in the boundary functions a_i and b_i. Therefore

$$\begin{aligned} Y_1 &= L_{11}Z_1, \\ Y_2 &= L_{21}Z_1 + L_{22}Z_2, \\ &\vdots \\ Y_p &= L_{p1}Z_1 + L_{p2}Z_2 + \ldots + L_{pp}Z_p. \end{aligned} \tag{3.9}$$

Assuming the constraints for variables $Y_1 \leq b_1$ and $Y_2 \leq b_2$, consequently it can be expressed by,

$$\begin{aligned} Z_1 &\leq L_{11}^{-1} b_1, \\ Z_2 &\leq \left(b_2 - L_{21}L_{11}^{-1}Y_1\right) L_{22}^{-1}. \end{aligned} \tag{3.10}$$

In general, the formula can be written for p-dimensional random variable at a multivariate truncated points b_i , $i = 1, ..., p$ as,

$$Z_i \leq L_{ii}^{-1}\left(b_i - \sum_{j=1}^{i-1} L_{ij}Z_j\right), \tag{3.11}$$

where $Z_j = L_{jj}^{-1}\left(Y_j - \sum_{k=j-i}^{i-2} L_{jk}Z_k\right)$. This algorithm can be extended for two-sided truncated multivariate normal distribution as,

$$L_{ii}^{-1}\left(a_i - \sum_{j=1}^{i-1} L_{ij}Z_j\right) \leq Z_i \leq L_{ii}^{-1}\left(b_i - \sum_{j=1}^{i-1} L_{ij}Z_j\right). \tag{3.12}$$

Cholesky Algorithm for TMVN

We here propose the generalized algorithm using matrix factorization,

- **Step 1:** Generate $u_1, ..., u_n$ from uniform distribution $u \sim U_{[0,1]}$.
- **Step 2:** Given a symmetric positive definite covariance matrix, $\Sigma_{p\times p}$, n, a, b.
- **Step 3:** Compute lower triangular Cholesky factor L from $\Sigma = LL^T$.
- **Step 4:** Boundary intervals can be presented as $[a_i, b_i]$ where truncated domain defined for $i = 1, ..., p$.
- **Step 5:** Set $Z_i = \Phi^{-1}(\alpha_i + u(\beta_i - \alpha_i))$ for $i = 1, 2, ..., p$.
  where $\alpha_i = \Phi\left(L_{ii}^{-1}(a_i - \sum_{j=1}^{i-1} L_{ij}Z_j)\right)$ and $\beta_i = \Phi\left(L_{ii}^{-1}(b_i - \sum_{j=1}^{i-1} L_{ij}Z_j)\right)$.
- **Step 6:** Obtain Z_j by $Z_j = L_{jj}^{-1}\left(Y_j - \sum_{k=j-i}^{i-2} L_{jk}Z_k\right)$.

3.4 General Linear Time Series

Let z_t, $t = 1, ..., n$, be an ergodic stationary Gaussian time series with mean μ and autocovariance (acvf) function $\gamma_k = Cov(z_t, z_{t-k})$, $k = 0, ..., n-1$. Suppose for a given observed series of size n, autocorrelation (acf) be specified by $\rho_k = \gamma_k/\gamma_0$, and the covariance matrix of z_t can be obtained by,

$$\Gamma_n = \gamma_{i-j}, \tag{3.13}$$

where the (i, j) entry is denoted in the $n \times n$ matrix. The general linear process (GLP) model can be written, in general, as

$$z_t = \mu + a_t + \Psi_1 a_{t-1} + \Psi_2 a_{t-2} + ..., \tag{3.14}$$

where a_t, $t = 1, 2, \ldots$ is a sequence of independent normal random variables with zero mean and σ_a^2 variance and $\Sigma_k \Psi^k < \infty$. A general ARMA(p,q) model is given by,

$$z_t = \phi_0 + \sum_{i=1}^{p} \phi_i z_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i}, \tag{3.15}$$

where $a_t \sim IID(0, \sigma_a^2)$. Let us assume that ARMA(p,q) model is of the form

$$\phi(B)(z_t - \mu) = \theta(B)a_t, \tag{3.16}$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$ where B is a back-shift operator on the index of the time series. ARMA(p,q) can be specialized as a pth-order autoregressive process and a moving average process of order q. Thus, z_t and z_{t-k} can be jointly distributed as bivariate normal distribution.
This model can also be defined as,

$$\phi(B)z_t = \zeta + \theta(B)a_t, \tag{3.17}$$

where ζ is the intercept parameter is $\zeta = \phi(1)\mu$. The essential requirement for stationary and invertibility is that all roots of the polynomial equation $\phi(B)\theta(B) = 0$ lie outside the unit circle where B is a complex variable in this equation. We will further proceed to examine the stationary and invertible autoregressive fractionally integrated moving average ARFIMA(p,d,q) where the model equation can be expressed in general as $\nabla^d \phi(B) z_t = \zeta + \theta(B)a_t$ for $\mid d \mid < 1/2$, and also $\phi(B)$ and $\theta(B)$ have the same properties as defined for ARMA process. In R software, a negative of the moving average coefficients is widely used, following by $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$.

In ltsa package, the Durbin-Levinson and Davies-Harte algorithms used to simulate General Linear Process in time series [McLeod et al., 2007]. In this section, we introduce the algorithm to generate Box-Cox time series by the given autocovariance function (acvf) $\gamma_k = Cov(z_t, z_{t-k})$ and the covariance matrix specified by Γ_k. Simulation of truncated normal distribution by employing Cholesky decomposition is exact method. We define the optimal one-step-ahead forecast of z_{n+1} that is given by,

$$\hat{z}_{n+1} = E(z_{n+1}|z_1, ..., z_n) = \mu + \phi_{n1}(z_n - \mu) + ... + \phi_{nn}(z_1 - \mu), \tag{3.18}$$

where

$$\Gamma_n(\phi_{n1}, .., \phi_{nn}) = \gamma_n. \tag{3.19}$$

Durbin-Levinson algorithm provides a fast solution to the linear prediction problem and it is used to generate general linear Gaussian time series with given covariance matrix [Brockwell and Davis, 1991]. Durbin-Levinson solves by only using $O(n^2)$ arithmetic operations as opposed to the $O(n^3)$ operations in Cholesky factorization.

3.5 Simulation of Box-Cox Time Series

In this section we mention some further applications and extensions of the Levinson algorithm. To simulate the Box-Cox time series, using the fact that Box-Cox and its inverse transformed data have truncated normal distribution. We design a simulation procedure based on generating initial values obtained by matrix factorization in order to avoid bias.

3.5.1 BoxCoxAR(1) Time Series Analysis

In fact, a simple time series model may be useful to illustrate the contribution of exact Box-Cox analysis in the field of time series. The simple autoregressive model, AR(1) be as follows,

$$z_t = \phi_0 + \phi_1 z_{t-1} + a_t, \tag{3.20}$$

where a_t assumed to be white noise series with mean zero and variance σ_a^2. Under the stationary condition, we can define z_t from Box-Cox distribution. Hence, the mean and the variance of z_t are given by,

$$\begin{aligned} E(z_t) &= \mu, \\ Var(z_t) &= \frac{\sigma_a^2}{1-\phi_1^2}. \end{aligned} \tag{3.21}$$

If we assumed that the latent series as z_1,..., z_n, then the joint probability density function can be written as,

$$f(z_1, ..., z_n) = f(z_1)f(z_2|z_1)f(z_3|z_2)...f(z_n|z_{n-1}). \tag{3.22}$$

Consequently, the conditional distribution $f(z_t|z_{t-1})$ for $t = 2, ..., n$ is normal with mean $\mu_t = \mu + \phi(z_{t-1} - \mu)$ and variance σ_a^2.
In the first step, we consider AR(1) model to simulate the Box-Cox transformed time series and its inverse in the original domain. Hence, we have

$$(z_t^{(\lambda)} - \mu) = \phi(z_{t-1}^{(\lambda)} - \mu) + a_t, \tag{3.23}$$

where $a_t \sim N(0, \sigma_a^2)$.
Assuming that $z_t^{(\lambda)}$ series are stationary, then we can denote $z_t^{(\lambda)} = w_t$ with non-zero mean. The initial value, w_1, obtained from $w_1 \sim TN(\mu, \frac{\sigma_a^2}{1-\phi_1^2}, a, b)$. Hipel and McLeod [1994] suggested to employ WASIM1 or WASIM2 algorithms to compute w_1 in order to eliminate bias. ϕ_1 is estimated by data or given. Next, we can generate w_t from the past time series w_{t-1} and a_t is simulated by computer.

We illustrate a simulated BoxCoxAR(1) model with left truncation point at $-\lambda^{-1}$ where $\lambda = 0.25, 0.5, 0.75, 1$. Then, time series of length $n = 500$ with $\mu = 0$ and $\sigma_a = 1$ and $\phi = 0.8$ is shown in Figure 3.3. Figure 3.4 displays a simulated BoxCoxAR(1) with $\mu = -10$ and various $\lambda < 0$. As can be seen in Table 3.1, the different values of the Box-Cox transformation lead to change in the forecasts and its standard deviations.

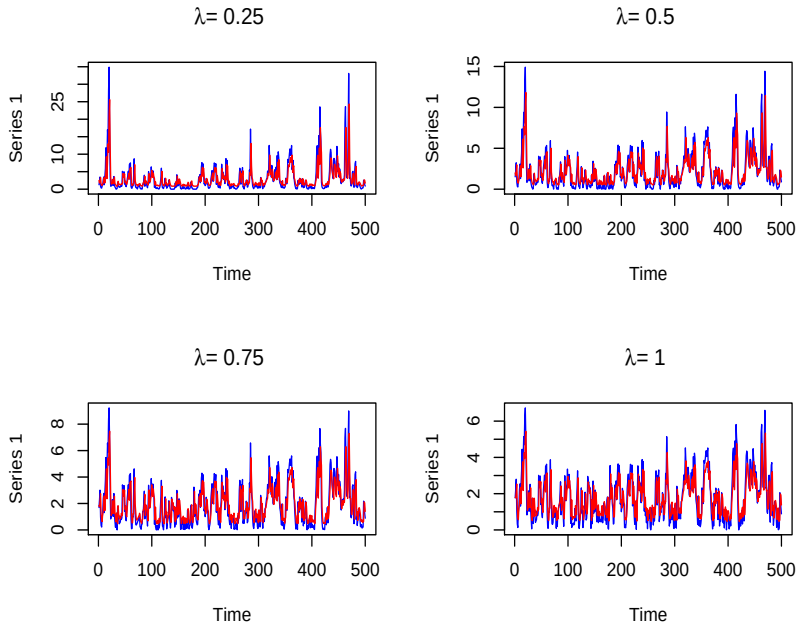

Figure 3.3: Comparison between the simulated BoxCoxAR(1) time series with different Box-Cox transformations $\lambda = 0.25, 0.5, 0.75, 1$.

	$\phi = 0.8$		$\phi = -0.8$	
lambda	Forecast	Standard deviation of forecast	Forecast	Standard deviation of forecast
0.25	1.494	2.936	2.623	2.984
0.5	1.214	1.515	2.176	1.698
0.75	1.139	1.076	1.870	1.338
1	1.123	0.877	1.653	1.119

Table 3.1: Forecasts and their standard deviations at lead time $l = 1$ for fitted AR(1) model to simulated time series.

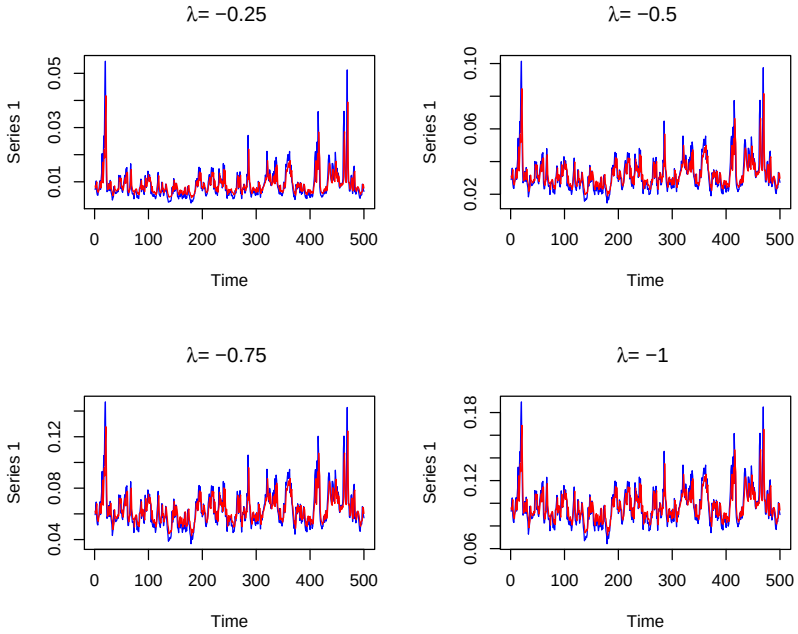

Figure 3.4: Comparison between the simulated BoxCoxAR(1) time series with different Box-Cox transformations $\lambda = -0.25, -0.5, -0.75, -1$.

3.5.2 Simulate Sunspot Time Series Model

The statistical model for yearly sunspot numbers from 1700 to 1988 are used to make comparison in the transformed and original domain. Using FitAR, we fit a constrained AR(9) to the sunpot numbers. Parameters estimates are given below. Approximate Box-Cox analysis produced $\hat{\lambda} = 0.463$. Since $k \doteq 99\%$, and an exact Box-Cox approach would gives virtually the same results.

The series was simulated using the rejection algorithm. Series of length $n = 10^4$ generated. The rejection algorithm was used 106 times, that is, about 1% of the time.

Box-Cox AR(9) Gaussian Simulation

We use the accept-reject algorithm as in Robert [1995] to simulate a realization with $n = 10^4$ observations from the subset AR(9) model that was fitted using our R script. We simulate a long time series so that we may make accurate comparisons with the expected theoretical autocorrelation and the sample values. Also we examine histogram of the marginal distribution of the data.

We found that the rejection algorithm was needed 118 times during the 10^4 simulations which agrees approximately with our expected value $1 - \kappa = 1.7\%$.

We can compare the original data with the simulated data in the original data domain as illustrated in Figures 3.5 and 3.6.

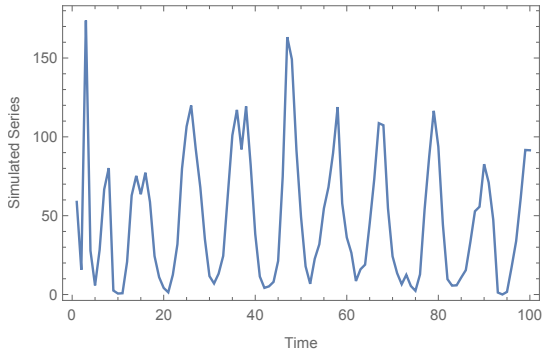

Figure 3.5: Simulated sunspot time series.

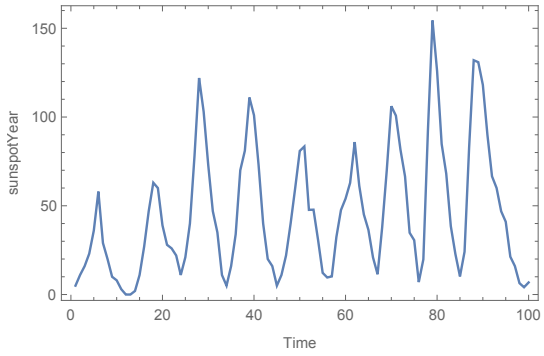

Figure 3.6: Yearly sunspot time series

We compare the theoretical and sample autocorrelations. As expected the agreement is satisfactory.

The theoretical specification for the autocorrelation in the original domain was derived by Granger and Newbold [1976] using Hermit polynomial expansion and it is quite complicated. Alternately, we can estimate the theoretical or expected autocorrelation by simulation. The autocorrelation function of this nonlinear time series can be compared with expected autocorrelations from the corresponding linear process.

The theoretical and sample autocorrelations of simulated Box-Cox Gaussian time series are discussed in Figures 3.7 and 3.8.

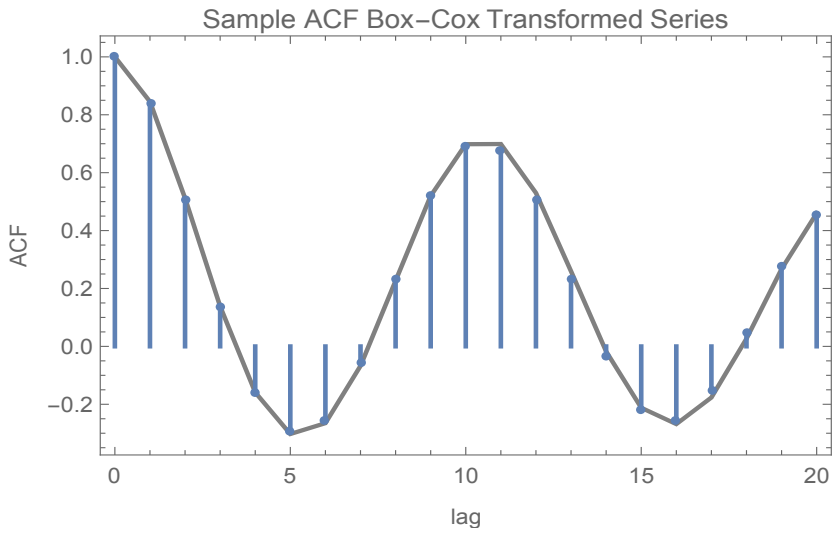

Figure 3.7: Theoretical and sample autocorrelations for simulated Box-Cox transformed time series.

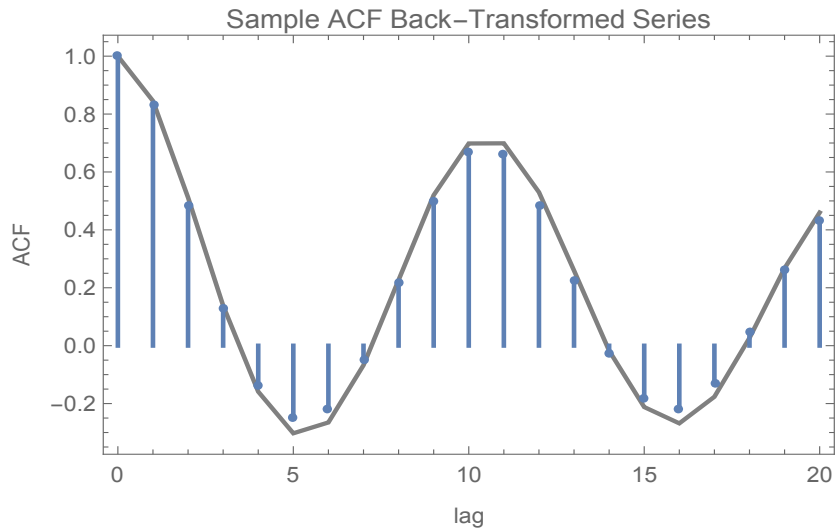

Figure 3.8: Theoretical and sample autocorrelations for simulated back-transformed time series.

Forecastability

Let $\{y_t\}$ be a covariance stationary time series with autocovariance function $\{\gamma_k\}$ and let $\Gamma_n = (\gamma_{i-j})_{n\times n}$ be the covariance matrix of n successive observations, $z_1, \ldots, z_n$. Then assuming there is no deterministic component, so $E\{y_t\} = 0$, then there exists a linear process so that we may write,

$$y_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots \tag{3.24}$$

where a_t are the white noise innovation with variance σ_a^2. An accurate approximation to the innovation variance is given by

$$\sigma_a^2 = \frac{\det(\Gamma_n)}{\det(\Gamma_{n-1})} \tag{3.25}$$

for n large enough. In the case of the sunspot AR(9) model, $n = 100$, is large enough to obtain a highly accurate approximation. The coefficient of forecastability [Granger and Newbold, 1976] may be written,

$$R^2 = 1 - \frac{\sigma_a^2}{\sigma_y^2} \tag{3.26}$$

where $\sigma_y^2 = \gamma_0$ is the variance of the time series.

We fit the subset AR(9) model and after Box-Cox analysis, we obtain the fitted model,

$$z_t = y_t^{(\hat{\lambda})} \tag{3.27}$$

and

$$w_t = z_t - \bar{z} \tag{3.28}$$

Hence, we have

$$w_t = 1.245 w_{t-1} - 0.527 w_{t-2} - \ldots - 0.202 w_{t-9} + \hat{a}_t \tag{3.29}$$

where $\hat{\sigma}_a = 2.060771$, $\bar{z} = 10.903$ and $\hat{\lambda} = 0.463$.

Using this fitted model we evaluated the coefficient of forecastability for the Box-Cox transformed model and obtained $R_z^2 = 88.8\%$.. Then we back-transformed the simulated series and computed the coefficient of forecastability in the original data domain to obtain $R_y^2 = 86.1\%$. This agrees with the findings of Granger and Newbold [1976]. In general the result of Granger and Newbold [1976] can be interpreted to say that if the correct model is found after a suitable Box-Cox transformation then its coefficient of forecastability is larger than a suitable model that could be fit in the original data domain.

As we can see from Table 3.2, both the empirical simulation and theoretical probability computation indicate about 90% of the inverse transformation is invalid. Empirically we find the true kappa based on 10^4 simulations shown in Table 3.3.

	λ	ξ	$1-\kappa$	simulation
sunspot.year	0.463	-2.307	0.011	0.872

Table 3.2: The theoretical probability of invalid back transform

	$\hat{\kappa}$	95% MOE
sunspot.year	0.983	0.003

Table 3.3: The true kappa based on 10,000 empirical simulations.

3.5.3 Exact Simulation of BoxCoxAR(p)

Simulation of AR and ARIMA models can be obtained in terms of non-Gaussian innovations from built-in R function. Initial values randomly computed from the specific model equation. McLeod [1975] pointed out to use an exact method for Gaussian time series in order to handle the length of burn-in period. A proposed algorithm can be employed to simulate the BoxCoxAR(p) for Gaussian process. Assuming that the initial time series values z_1,..., z_p are obtained from multivariate truncated normal distribution with mean (μ, ..., μ) and covariance matrix $\Gamma_k = (\gamma_{i-j})_{p\times p}$ by using Matrix factorization. Then, we can compute the remaining values from the model equation recursively for $t = p+1, ..., n$ as,

$$z_t - \mu = \phi_1(z_{t-1} - \mu) + ... + \phi_p(z_{t-p} - \mu). \tag{3.30}$$

This simulation approach is based on given autocovariance and autocorrelation function for Box-Cox transformed Gaussian time series. The following algorithm presents the exact simulation of stationary AR(p) models.

- **Step 1:** Given autocovariance function (acvf) $\gamma_k = Cov(z_t, z_{t-k})$, $k = 0, ..., p-1$ or any desired covariance matrix, Γ_k.

- **Step 2:** Determine the lower triangular matrix L by Cholesky decomposition for $\Gamma_k = LL^T$.
- **Step 3:** $z_1, ..., z_t$, $t = 1, ..., p$, can be generated from the exact Box-Cox Normal Distribution.
- **Step 4:** Obtaine $y_1, ..., y_p$ using eqn. (3.32) in the transformed domain.
- **Step 5:** Simulate the transformed time series from eqn. (3.30), and then the Box-CoxAR(p) model is computed in the untransformed domain.

$$\Gamma_k = LL^T = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix}, \tag{3.31}$$

and

$$Y_t = LZ_t. \tag{3.32}$$

3.5.4 Numerical Example

Figure 3.9 illustrates the time series plot for Ninemile series. The Ninemile time series consists of $n = 771$ observations of the annual treeing width measurement on Douglas fir at Nine Mile Canyon, Utah from 1194 to 1964. We will employ the Box-Cox transformation in order to improve a statistical model.

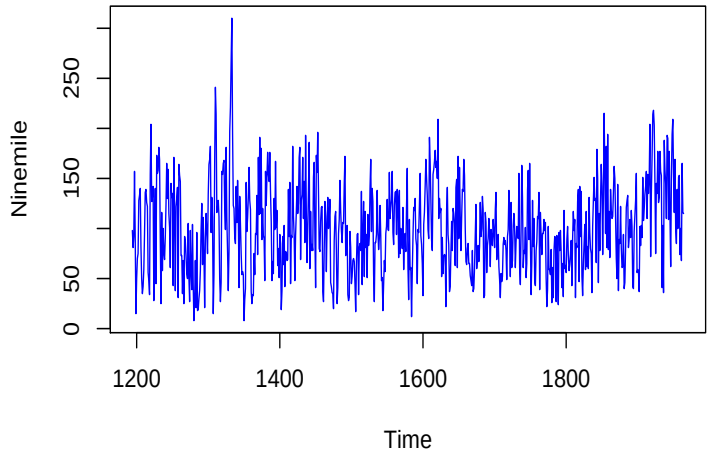

Figure 3.9: Time series plot of Ninemile time series

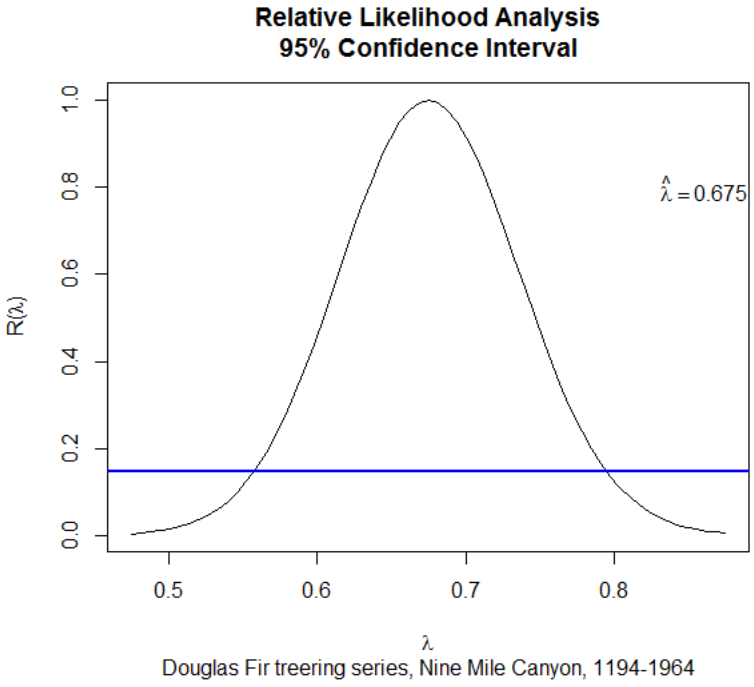

Figure 3.10: Box-Cox analysis produced by BoxCox(Ninemile) for fitted AR(1).

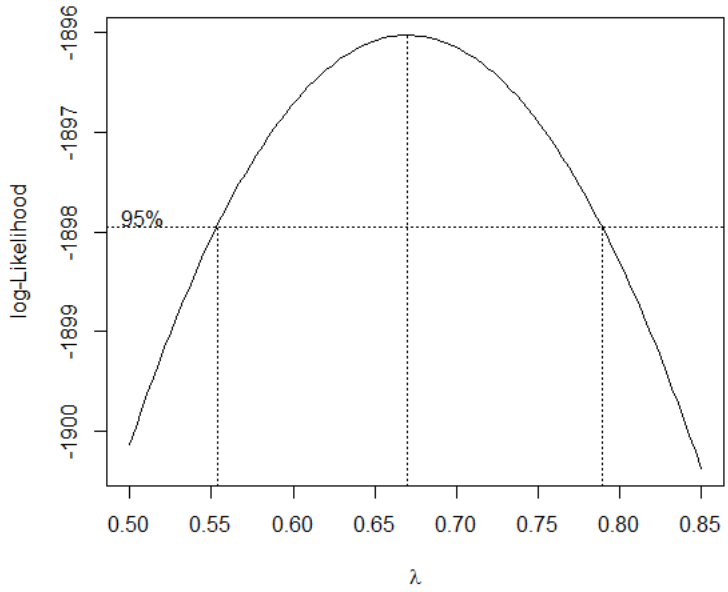

Figure 3.11: Graph from boxcox for fitting ARp(1, 2, 6, 9) to Ninemile series.

	λ	κ	failure probability
Ninemile	0.675	0.999246	0.075%

Table 3.4: The κ and the probability of the Box-Cox normal approximation is failed shown for Ninemile time series.

Figure 3.10 displays the log-likelihood as a function of the power parameter, λ. The maximum occurs at $\lambda = 0.675$, but a power transformation with $\lambda = 0.5$ is not within the confidence interval for λ. We will take Box-Cox optimal transformation of the Ninemile values for further analysis. The results in Table 3.4 show that $\kappa \doteq 1$, hence there is very slightly failure for each of the 10000 simulations. It can be used a built-in function SelectModel to determine the best ARp model. For the Ninemile time series, we fit the ARp(1, 2, 6, 9) model by least-squares using FitARp. In this case, the output from FitARp obtained and the MASS boxcox() built-in R function employs approximate log-likelihood corresponding to linear regression as illustrated in Figure 3.11. The difference in Figure 3.10 and 3.11 is not significant. As we can see, both plots strongly suggested the approximately same optimal transformation.

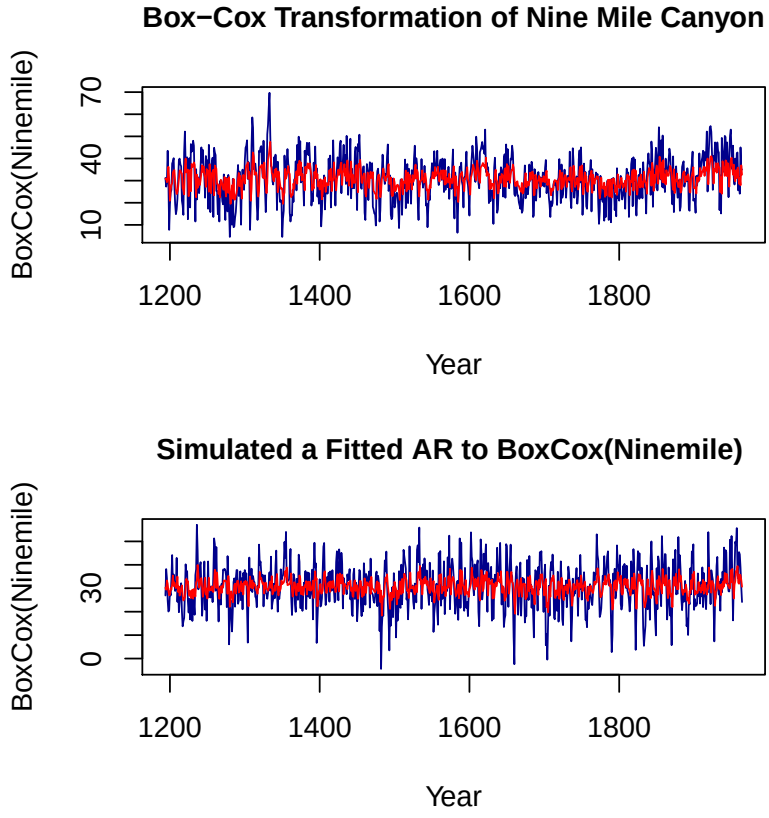

Figure 3.12: Box-Cox transformed of Ninemile in transformed scale illustrated in first panel, and also simulation of transformed Ninemile series from fitted AR(1) model via bootstrap method as shown in the second panel.

Figure 3.12 illustrates BoxCox(Ninemile) and simulated a fitted AR(1) to Box-Cox transformed Ninemile, $\hat{\lambda} = 0.675$. In fact, the parameter estimates in Ninemile series are used to simulate time series from AR(1) model. We generate a time series from BoxCoxAR(1) model in transformed domain, and then fitted the AR(1) model to the Box-Cox transformed simulated data. We examine the estimated residuals and try to evaluate power transformation and determine the 95% confidence interval of λ by bootstrap simulation. In Figure 3.13, the transformed actual Ninemile and simulated series in transformed scale are shown.

To compare the performance for some measures of forecast accuracy, we see from Table 3.5 that accuracy measures for simulated series are very close to the Box-Cox transformed Ninemile. If we compare the log-likelihood for Ninemile data with simulated data after fitting AR(1) model, it seems that similar results can be obtained.

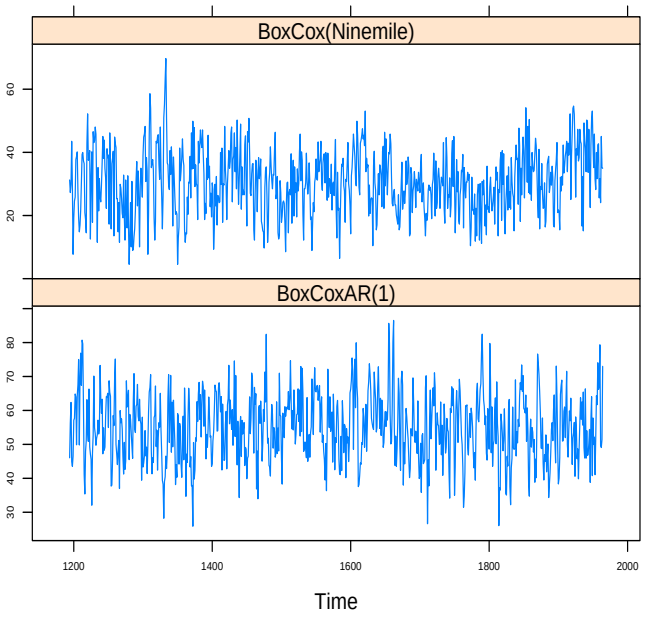

Figure 3.13: Comparison of Box-Cox transformed Ninemile series and simulated Box-Cox transformed in the transformed data domain.

Accuracy measure	BoxCoxAR(1)	BoxCox(Ninemile)
RMSE	8.98	9.17
ME	0.006	0.002
MAE	7.23	7.44
MASE	0.85	0.86

Table 3.5: Different accuracy measures for simulated Box-Cox transformed AR(1) series and Box-Cox transformed Ninemile.

3.6 Modified D-L Algorithm for Box-Cox Time Series

This section focuses on the simulation of BoxCoxARMA model using the Durbin-Levinson procedure. Brockwell and Davis [1991] proposed Durbin-Levinson approach to generate Gaussian time series with a given covariance structure. We first describe an algorithm of generating a linear time series proposed by McLeod et al. [2007, 2012] based on the Durbin-Levinson algorithm [Brockwell and Davis, 1991]. The recursive algorithm that is discussed in this section is applicable to determine the best linear predictor of z_{k+h} in terms of z_k,..., z_1. Define v_k as the variance of the k step linear predictor. Let,

$$\phi_{1,1} = \gamma_1/\gamma_0, \tag{3.33}$$

and

$$v_1 = (1 - \phi_{1,1}^2)v_0, \tag{3.34}$$

where $\gamma_0 = v_0$. Then we can recursively obtain the coefficients $\phi_{k,1}$, ...,$\phi_{k,k}$ from the equations,

$$\phi_{k,k} = \left[\gamma_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\gamma(k-j)\right]/v_{k-1}, \tag{3.35}$$

$$\begin{bmatrix} \phi_{k,1} \\ \vdots \\ \phi_{k,k-1} \end{bmatrix} = \begin{bmatrix} \phi_{k-1,1} \\ \vdots \\ \phi_{k-1,k-1} \end{bmatrix} - \phi_{k,k} \begin{bmatrix} \phi_{k-1,k-1} \\ \vdots \\ \phi_{k-1,1} \end{bmatrix}. \tag{3.36}$$

and

$$v_k = (1 - \phi_{k,k}^2)v_{k-1}. \tag{3.37}$$

Suppose z_t is a zero mean stationary time series with autocovariance function $\gamma(.)$, then we can predict the h step-ahead predictor based on the previous observations as follows,

$$\hat{z}_{k+1} = \phi_{k,1}z_k + ... + \phi_{k,k}z_1. \tag{3.38}$$

An equivalent formulation for ARMA(p,q) with given autocovariance function, γ_0, ..., γ_{k-1} for $k = 1, ... , n$ is defined by,

$$z_k = \phi_{k-1,1} z_{k-1} + ... + \phi_{k-1,k-1} z_1 + a_k. \tag{3.39}$$

Assuming the initial value in Durbin-Levinson generated by truncated normal distribution with the truncation point depended on $T = -1/\lambda$. Then, we can show that z_1, z_2, ... , z_n are recursively simulated,

$$\begin{aligned} z_1 &\sim TN_{(\lambda)}(\mu, \gamma_0), \\ z_2 &= \phi_{1,1} z_1 + a_2, \\ z_3 &= \phi_{2,1} z_2 + \phi_{2,2} z_1 + a_3, \\ &\vdots \\ z_n &= \phi_{n-1,1} z_{n-1} + ... + \phi_{n-1,n-1} z_1 + a_n. \end{aligned} \tag{3.40}$$

where $\sigma_z^2 = \gamma_0$. This model can be written, in general, as

$$\begin{aligned} \mu_n &= \phi_{n-1,1} z_{n-1} + ... + \phi_{n-1,n-1} z_1, \\ z_n &\sim TN_{(\lambda)}(\mu_n, \sigma_a^2). \end{aligned} \tag{3.41}$$

where $n = t$.

Simulation Algorithm

- **Step 1:** Given autocovariance function γ_k for $k = 0, 1, ..., n-1$ by using with σ_a^2.
- **Step 2:** Using Cholesky decomposition, $\Gamma_k = LL^T$, to compute a lower triangular matrix.
- **Step 3:** Simulate e_1, ..., e_p from the Box-Cox Normal distribution with none-zero-mean and σ_a^2 variance.
- **Step 4:** Generate initial value, z_1, ... , z_p by using $z_t = \sum_{k=1}^{t} L_{t,k} e_k$ for $t = 1, ..., p$.
- **Step 5:** Obtain z_{p+1}, ... , z_n time series, using Durbin-Levinson algorithm as,

$$z_t = \phi_{t-1,1} z_{t-1} + ... + \phi_{t-1,t-1} z_1 + a_t, \qquad t = p+1, 2, ..., n. \tag{3.42}$$

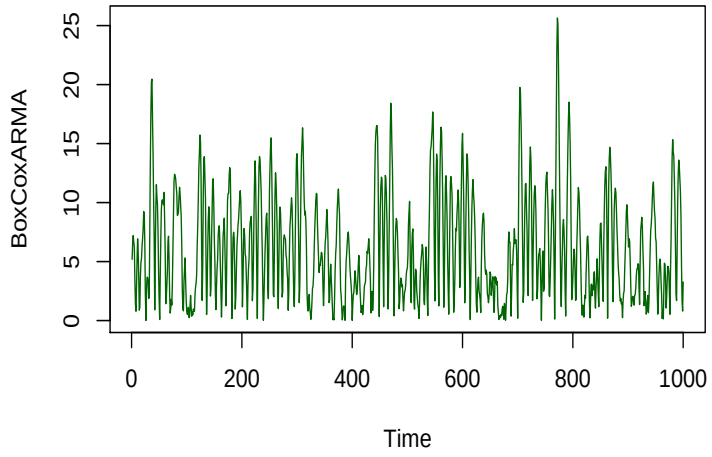

Figure 3.14: Simulate the BoxCoxARMA time series with $\lambda = 1$.

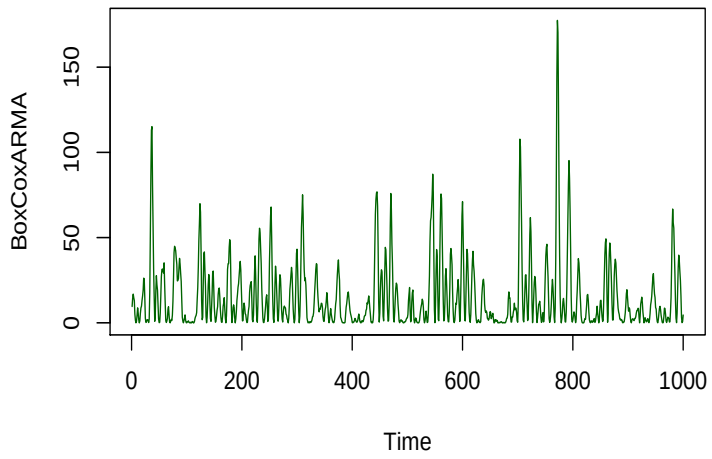

Figure 3.15: Simulate the BoxCoxARMA time series with $\lambda = 0.5$.

3.7 Appendix. Simulation of Time Series

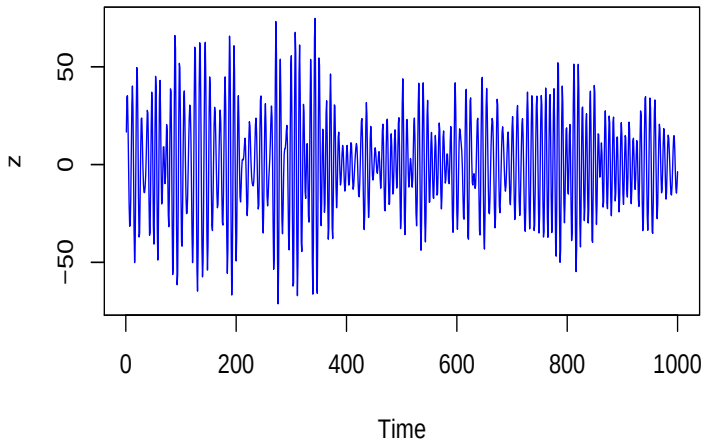

Figure 3.16: Simulate the transformed GuassianBoxCoxAR series with $\lambda = 0.5$.

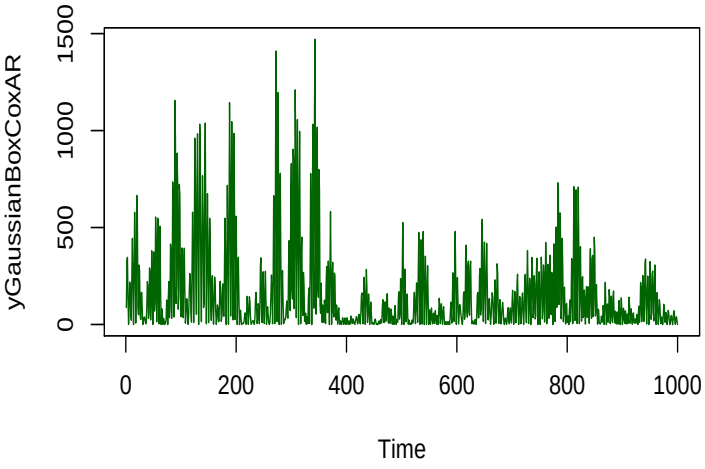

Figure 3.17: Simulate the GuassianBoxCoxAR time series with $\lambda = 0.5$ in the original domain.

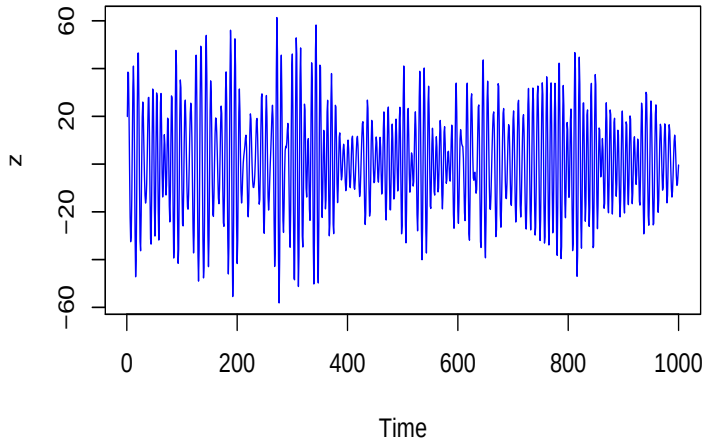

Figure 3.18: Simulate the transformed GuassianBoxCoxAR series with $\lambda = 1$.

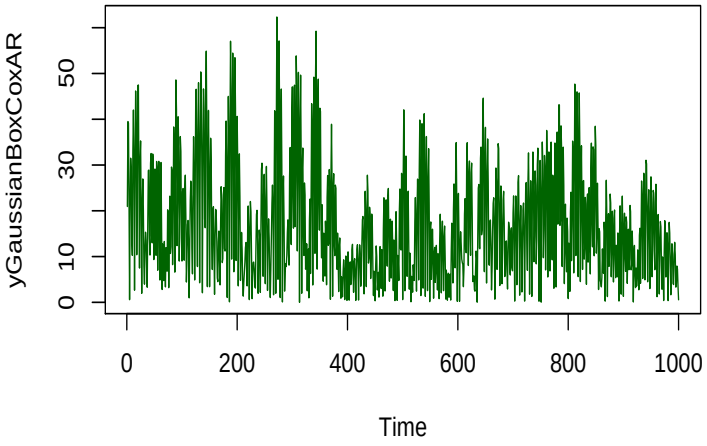

Figure 3.19: Simulate the GuassianBoxCoxAR time series with $\lambda = 1$ in the original domain.

Chapter 4

Conclusion

4.1 Summary of Transformations and Machine Learning

The main focus of machine learning (ML) is prediction. Box-Cox analysis provides a parametric approach to embed a Gaussian statistical model in a larger non-linear model which includes a continuous family of power transformations. It was demonstrated by Box and Cox [1964] that this method simplified the model by removing skewness as well as nonlinearities including interactions and heteroskedasticity. This suggests that perhaps a transformation may improve predictions from an ML model such as random forest (RF) when the output variable is skewed. Since we will not be using likelihood method to estimate the transformation, we consider the simple family of power transformations,

$$Y^{(p)} = \begin{cases} Y^p, & p \neq 0, \\ \log(Y), & p = 0. \end{cases} \tag{4.1}$$

and its inverse,

$$Y = \begin{cases} (Y^{(p)})^{1/p}, & p \neq 0, \\ \exp(Y^{(p)}), & \lambda = 0. \end{cases} \tag{4.2}$$

This power transformation is recommended when the output variable, y, has a skewed distribution and $y > 0$. In this situation, the usual mean-square error criterion

$$MSE(Y, \hat{Y}) = E\left[(Y - \hat{Y})^2\right], \tag{4.3}$$

is less appropriate. When the output is not close to zero, $y > 0$, the mean absolute percentage error

$$MAPE(Y, \hat{Y}) = E\left[\frac{|Y - \hat{Y}|}{Y}\right], \tag{4.4}$$

is often used. In practice, the MSE and MAPE are estimated by using sample averages. Other criteria may also be considered depending on the underlying problem. The power transformation may be carried in the cross-validation stage by evaluation possible values of p on a grid and selecting the transformation that provides the most accurate MAPE prediction in the original data domain. The steps are outlined as follows,

- **Step 1:** Verify that output variable is skewed and that when the ML algorithm applied to the data, the residuals are also skewed.
- **Step 2:** Select values of p. For example we may take $p = 1, 1/2, 1/3, 0, -1/2$.
- **Step 3:** Transform the output variable using eqn. (4.1) and apply the ML algorithm to obtain predictions in the transformed domain.
- **Step 4:** Apply the inverse transformation in eqn. (4.2) to obtain predictions in the original data domain.
- **Step 5:** Evaluate the average prediction error (EPE) using a suitable criterion.
- **Step 6:** Repeat Step 3-5 for each p and select the best p which provides the best predictions.

After the optimal p is found on the training data, the training EPE is compared with the test EPE to check for overfitting.

4.1.1 Application to Boston Housing dataset

The Boston Housing dataset is provided available in R (Boston::MASS) as well as in the curated datasets in Mathematica.

The original researchers were interested in explaining what factors were important in determining the median value of owner-occupied houses in the city and suburbs around Boston. Average aggregate data were obtained for n = 506 Boston suburbs for 14 variables. The boxplot of the output variable, median value of owner-occupied house, is skewed to the right as illustrated in Figure 4.1.

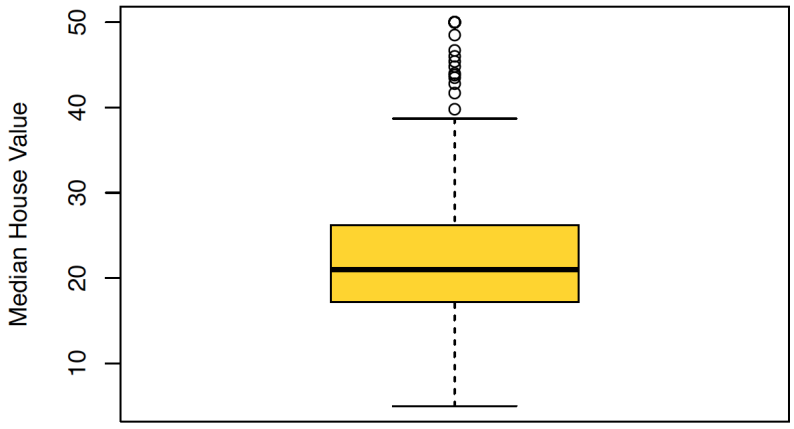

Figure 4.1: Median House Price, Training set shown.

Figure 4.2 demonstrates that the residuals in both the training and test samples are positive skewed as might be expected from Figure 4.1. Hence, a transformation might improve the accuracy of the predictions even in the original data domain.

Figure 4.2: Boxplot of the Training and Test residuals for random forest.

Comparing random forest with multiple linear regression on the training and test dataset it was found that random forest usually performs much better than linear regression. We see from Table 4.1 that RF seems to outperform Linear Regression even with no transformation. Table 4.1 compares the RMSE for linear regression and RF using an average of 10^2 iterations.

	Training Data	Test Data
Linear Regression	4.4	5.8
Random Forest	2.4	4.1

Table 4.1: RMSE comparison using average of 100 replications.

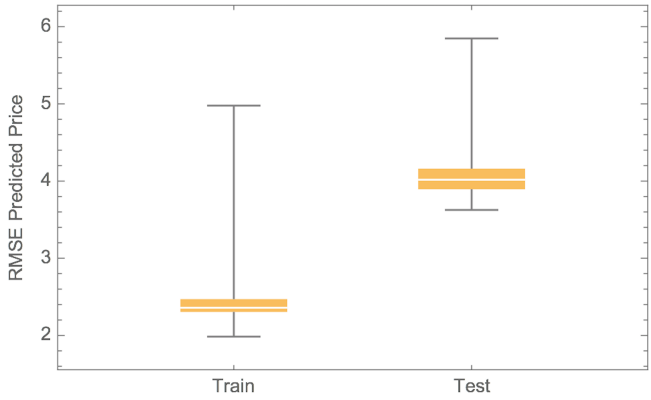

Figure 4.3: RMSE shown for random forest based on 100 replications.

As can be seen from Figure 4.3 and Table 4.1 the interquartile range is reasonably narrow so about half the time the predictions are very good. However, Figure 4.3 shows that there

is a long-right tail to the output from the RF method producing RMSE's that are very large indicating poor quality predictions for that iteration. This is a serious drawback not shared by more deterministic algorithms such as linear regression, artificial neural nets and support vector machines. As we can see from Table 4.2, RF is better than regression. The predictions for RF were based on 100 replications of the RF and averaging all 100 predictions.

	Training Data	Test Data
Linear Regression	16.0	17.6
Random Forest	7.9	11.5

Table 4.2: MAPE comparison using average of 100 replications for linear regression and random forest.

RF predictions based on 1000 replications for various power transformations are shown in Table 4.3. Table 4.3 reveals that $p = -0.5$ produces the smallest average MAPE in the training data. For simplicity a log transformation can be used and the expected improvement in MAPE is about 7.6%. This improvement is also statistical significant as shown by 95% MOE given in Table 4.4.

p	Training Data	Test Data
1.	7.98	11.58
0.5	6.59	10.97
0.25	6.37	10.80
0.	6.25	10.70
-0.25	6.17	10.69
-0.5	6.14	10.73
-1.	6.31	10.91

Table 4.3: MAPE various power transformation for random forest.

p	Training Data	Test Data
1.	0.553	0.666
0.5	0.503	0.649
0.25	0.494	0.644
0.	0.490	0.641
-0.25	0.486	0.640
-0.5	0.485	0.642
-1.	0.492	0.647

Table 4.4: 95% MOE for estimates shown in Table 4.3.

4.2 Concluding Remarks

This thesis demonstrates that in some situations that exact Box-Cox likelihood analysis may provide a better approach than is currently used. Research to develop a more efficient and specialized optimization algorithm to obtain the exact profile log-likelihood is needed. To this end, we investigated the EM algorithm but were not able to obtain satisfactory convergence with this method for practical use. The work on the EM algorithm is summarized in Chapter 2.

More extensive simulations are also warranted to investigate the overall performance of the exact Box-Cox maximum likelihood estimates. It was pointed out that asymptotic theory is to provide insufficient guidance for all situations since the assumption that the term $-n \log \kappa$ can be neglected asymptotically is not generally valid. Bootstrapping and Monte-Carlo statistical tests provide a better alternative for statistical inference on the parameter λ and here it is helpful to use our exact approach to random sampling from the exact Box-Cox Data Distribution. As is shown in Chapter 3, simple rejection algorithms are impractical when Box-Cox methods are used with time series models.

It is evident that the simulation experience results are not affected by the exact Box-Cox Data Distribution in the case κ is close to 1. Further, we presented that in forecasting and simulation, truncation is likely significant, especially by using bootstrapping and cross-validation. As a result, the probability of failure may lead to loss the accuracy in analysis and forecasting. Regarding forecasting's concept, it seems that the reasonable unbiased forecast may be involved by using the correct distribution and the specific loss function.

In Chapter 3, we provide an efficient and direct algorithm via matrix factorization to generate the random variables from truncated normal distribution. The proposed simulation algorithm is more generalized for any covariance structure compared to Robert [1995]. Indeed, an iterative algorithm may deal with computational difficulty of convergence for multidimensional simulations. The fact that rejection sampling can be delicated to the probability of acceptance. With regard to time series models, the dependence assumption means that the computational demands of rejection method are even more sensitive to κ.

Later, the simulation approach is illustrated with an application to the Box-Cox time series. Moreover, the modified Durbin-Levinson recursions are extended to simulate a stationary Box-Cox time series given any desired autocovariance function. It is also shown that the BoxCoxAR and BoxCoxARMA models would be affected by λ and also the location parameter varies as the Box-Cox transformation changes.

Finally, we apply the power transformation for random forest and discover the improvement in the accuracy of the prediction when cross validation used to estimate the optimal transformation. Predictive performance of other ML algorithms such as MARS, SVM, MLP regression may sometimes be improved using power transformation.

In future, we will investigate the optimal forecast for the Box-Cox time series and any desired loss function.

Bibliography

E. B. Andersen. Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32(2):283–301, 1970.

A. C. Atkinson. Testing transformations to normality. *Journal of the Royal Statistical Society, Series B*, 35(3):473–479, 1973.

A. C. Atkinson and L. R. Pericchi. Grouped likelihood for the shifted power transformation. *Journal of the Royal Statistical Society, Series B*, 53(2):473–482, 1991.

M. S. Bartlett. The use of transformations. *Biometrics*, 3(1):39–52, 1947.

P. J. Bickel and K. A. Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311, 1981.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211–252, 1964.

G. E. P. Box and P. W. Tidwell. Transformation of the independent variables. *Technometrics*, 4(4):531–550, 1962.

G.E.P. Box, G. M. Jenkins, and G. C Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley, New York, 4th edition, 2008.

L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):614–619, 1985.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer, New York, 2nd edition, 1991.

R. J. Carroll. A robust method for testing transformations to achieve approximate normality. *Journal of the Royal Statistical Society, Series B*, 42(1):71–78, 1980.

R. J. Carroll and D. Ruppert. On prediction and the power transformation family. *Biometrika*, 68(3):609–615, 1981.

G. Casella and R. L. Berger. *Statistical inference*. Thomson Learning, 2nd edition, 2002.

G. Chen and R. A. Lockhart. Box-Cox transformed linear models: A parameter-based asymptotic approach. *The Canadian Journal of Statistics*, 25:517–529, 1997.

G. Chen, R. A. Lockhart, and M. A. Stephens. Box-Cox transformations in linear models: large sample theory and tests of normality. *The Canadian Journal of Statistics*, 30(2):177–209, 2002.

M. H. Chen and J. Deely. Application of a new Gibbs Hit-and-Run sampler to a constrained linear multiple regression problem. Technical report, Purdue University, 1992.

M. H. Chen and B. W. Schmeiser. General Hit-and-Run Monte Carlo sampling for evaluating multidimensional integrals. *Operations Research Letters*, 19:161–169, 1996.

R. C. H. Cheng and L. Traylor. Non-regular maximum likelihood problems. *Journal of the Royal Statistical Society, Series B*, 57(1):3–44, 1995.

N. Chopin. Fast simulation of truncated Gaussian distributions. *Statistics and Computing*, 21 (2):275–288, 2011.

A. C. Cohen. On estimating the mean and standard deviation of truncated normal distributions. *Journal of the American Statistical Association*, 44:518–525, 1949.

A. C. Cohen. Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *The Annals of Mathematical Statistics*, 21(4):557–569, 1950.

A. C. Cohen. *Truncated and Censored Samples: Theory and Applications*. CRC Press, 1991.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall/CRC, 1st edition, 1979.

P. Damien and S. G. Walker. Sampling truncated normal, beta and gamma densities. *Journal of Computational and Graphical Statistics*, 10(2):206–215, 2001.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

N. R. Draper and D. R. Cox. On distributions and their transformation to normality. *Journal of the Royal Statistical Society, Series B*, 31(3):472–476, 1969.

N. Duan. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610, 1983.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222:309–368, 1922.

A. M. Foster, L. Tian, and L. J. Wei. Estimation for the Box-Cox transformation model without assuming parametric error distribution. *Journal of the American Statistical Association*, 96 (455):1097–1101, 2001.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

A. E. Gelfand, A. F. M. Smith, and T. Lee. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87(418):523–532, 1992.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

C. W. J. Granger and P. Newbold. Forecasting transformed series. *Journal of the Royal Statistical Society, Series B*, 38(2):189–203, 1976.

D. A. Griffith. Better articulating normal curve theory for introductory mathematical statistics students: Power transformations and their back-transformations. *The American Statistician*, 67(3):157–169, 2013.

A. K. Han. A non-parametric analysis of transformations. *Journal of Econometrics*, 35:191–209, 1987.

D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

D. V. Hinkley and G. Runger. The analysis of transformed data. *Journal of the American Statistical Association*, 79(386):302–309, 1984.

K. W. Hipel and A. I. McLeod. *Time Series Modelling of Water Resources and Environmental Systems*. Elsevier, 1994.

N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions, 2nd Ed.* Boston : Houghton Mifflin, 1970.

J. D. Kalbfleisch and D. A Sprott. Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society, Series B*, 32(2):175–208, 1970.

J. D. Kalbfleisch and D. A. Sprott. Marginal and conditional likelihoods. *The Indian Journal of Statistics, Series A*, 35(3):311–328, 1973.

G. Lee and C. Scott. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56:2816–2829, 2012.

P. Li, J. Chen, and P. Marriott. Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96(2):411–426, 2009.

S. Liu, H. Wu, and W. Q. Meeker. Understanding and addressing the unbounded likelihood problem. *The American Statistician*, 69(3):191–200, 2015.

G. Marsaglia. Generating a variable from the tail of the normal distribution. *Technometrics*, 6: 101–102, 1964.

G. Marsaglia and W. Tsang. The ziggurat method for generating random variables. *Journal of Statistical Software*, 5(8), 2000.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 2nd edition, 2007.

A. I. McLeod. Derivation of the theoretical autocovariance function of autoregressive moving average time series. *Applied Statistics*, 24(2):255–256, 1975.

A. I. McLeod. *Symmetrizing Positive Random Variables*. Wolfram Demonstrations Project, 2009.

A .I . McLeod, Hao Yu, and Z. L. Krougly. Algorithms for linear time series analysis: With R package. *Journal of Statistical Software*, 23(5), 2007.

A. I. McLeod, H. Yu, and E. Mahdi. Time series analysis with R. *Time Series Analysis: Methods and Applications*, 30:661–701, 2012.

J. A. Montoya, E. Díaz-Francés, and D. A. Sprott. On a criticism of the profile likelihood function. *Statistical papers*, 50:195–202, 2009.

J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Society for Industrial and Applied Mathematics, 2000.

K. Pearson and A. Lee. On the generalised probable error in multiple normal correlation. *Biometrika*, 6(1):59–68, 1908.

D. J. Poirier. The use of the Box-Cox transformation in limited dependent variable models. *Journal of the American Statistical Association*, 73(362):284–287, 1978.

T. Proietti and M. Riani. Transformations and seasonal adjustment. *Journal of Time Series Analysis*, 30(1):47–69, 2009.

C. P. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125, 1995.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.

G. Rodriguez-Yam, R. A. Davis, and Louis L. Scharf. Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. unpublished manuscript, 2004.

J. Shao. *Mathematical Statistics*. Springer, 1998.

D. A. Sprott. *Statistical Inference in Science*. Springer New York, 2000.

J. M. G Taylor. The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81(393):114–118, 1986.

J. W. Tukey. On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28(3):602–632, 1957.

A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20(4):595–601, 1949.

C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11 (1):95–103, 1983.

I. K. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

Curriculum Vitae

Name:	Samira Soleymani
Post-Secondary Education and Degrees:	University of Western Ontario London, ON, Canada 2013 - 2018 Ph.D. (Statistics) Dalhousie University Halifax, NS, Canada 2010 - 2012 Ph.D. (Industrial Engineering)(Inc) Dalhousie University Halifax, NS, Canada 2008 - 2010 M.Sc. (Engineering Mathematics) Shahid Beheshti University Tehran, Iran 2001 - 2006 B.Sc. (Pure Mathematics)
Related Work Experience:	Teaching Assistant and Research Assistant The University of Western Ontario 2013 - 2018 Statistical Consultant The University of Western Ontario 2015 - 2017 Teaching Assistant and Research Assistant The Dalhousie University 2008 - 2012

Honours and Awards:

Faculty of Graduate Student (FGS) scholarship
The Dalhousie University
2011 - 2012

National Building Code of Canada (NBCC)
2009 - 2010

Conference Presentations:

- Samira Soleymani & A. Ian McLeod (2017)"Box-Cox Time Series", Computational Methods section, *SSC, The 45th Annual Meeting of the Statistical Society of Canada, Winnipeg, Canada*.
- Samira Soleymani & A. Ian McLeod (2016)"Simulation of Box-Cox Tranformed Time Series", Computational Methods section, *SSC, The 44th Annual Meeting of the Statistical Society of Canada, St. Catharines, Ontario, Canada*.
- Yuanhao Lai, Jianpei Jiang, & Samira Soleymani (2016)"Predicting the Number of the Test for Influenza and other Respiratory illnesses from Google Flue Trends", *SSC, The 44th Annual Meeting of the Statistical Society of Canada, St. Catharines, Ontario, Canada*.
- Samira Soleymani & A. Ian McLeod (2015)"Optimal Box-Cox Kriging", *Fallona Family Interdisciplinary Showcase Poster, Western University, London, Canada*.